STUDIES ON PERFORMANCE EVALUATION OF INTEGRATED COMMUNICATION SYSTEMS(Dissertation_全文)

Ohtsuki, Kazuhiro

②

Studies on Performance Evaluation of Integrated Communication Systems

KAZUHIRO OHTSUKI

Preface

The research reported in this dissertation deals with integrated communication systems. Integrated communication systems accommodate all kinds of existing services such as telephone, facsimile and data transfer. They are also conceived to provide various future services such as video, broadcast communication, etc. By the sharing of transmission and switching facilities, integration of a wide variety of applications in communication systems will provide high system availability in planning, development, installation and operation.

The fundamental problem of integration arises from the differences among various kinds of services both in traffic property and in transmission demand. This has resulted in new problem areas in communication networks design and in their performance evaluations as well.

The major objective of this research is to develop new transmission and switching techniques in integrated communication systems and to evaluate their performances. We focus on the packet switching technique, since the technique is considered as the most powerful means of integration. We roughly classify the integrated networks by two categories; one is the alternative in which network uses the existing network or constructs a new high-speed facilities, and another is the geometrical ranges of networks.

In Chapter 2, we consider a wide area network with the use of store-and-forward packet switching technique. Store-and-forward packet switching technique is the most popular switching technique in the conventional data communication networks. The network is modeled as a type of tandem queueing system with voice and data packets, where a voice packet is lost when the buffer of the switching node is full. We approximately analyze the end-to-end transmission delay and loss probability of voice packets.

Next, we consider the integration onto local area networks (LANs). We propose, in Chapter 3, the effective hybrid protocol on a CSMA/CD LAN, and analyze the throughput and delay for connectionless packets and the loss probability of the call for connection oriented packets. In Chapter 4, we evaluate the performance of a token passing ring network through extensive simulations to explore the feasibility of this class of networks as a medium for combined voice and data transmission. In Chapter 5, we introduce a type of bulk service queueing system which models the future integrated communication systems. The transmission discipline considered in this chapter is applicable to token-based access method such as token passing ring, token passing bus and FDDI.

Finally, we develop a high-speed integrated systems for future wide area networks. In Chapter 6, we propose a high-speed packet switch architecture based on multichannel allocation policy. We also evaluate the performance of the proposed switch by approximation analysis and simulations.

The research of the integrated communication systems is still in progress. The author would like to hope that this research will be helpful for further study in this field.

January, 1991

Kazuhiro Ohtsuki

ACKNOWLEDGEMENT

I wish to express my sincere appreciation to Professor *Toshiharu Hasegawa* of Kyoto University for his invaluable guidance and helpful suggestions since my student days and for supervising this dissertation. His constant encouragement and comments have greatly helped me accomplish this research.

I am very grateful to Associate Professor *Yutaka Takahashi* for his enthusiastic guidance and persistent encouragement. Discussions with him led me to make further research in the field of performance evaluation on computer communication systems.

I am highly indebted to Associate Professors *Hiromi Okada* of Osaka University and *James F. Kurose* of University of Massachusetts for their invaluable comments and guidance, in particular, concerning the work in Chapter 6. I am also grateful to Associate Professor *Tatsuya Suda* of University of California at Irvine for his earnest discussions and guidance for the work in Chapter 4.

I wish to acknowledge to Professors *Toshihide Ibaraki* and *Satoshi Hoshino* of Kyoto University for their invaluable comments and criticisms.

I would like to thank Associate Professor *Koichi Mutsuura* of Takuma National College of Technology, Associate Professor *Shin-ichi Ueshima* of Kansai University, Associate Professor *Kunio Goto* of Nanzan University, Mr. *Toshio Shimizu* of Matsushita Electric Industrial Co. Ltd., and Assistant Professor *Tetsuya Takine* of Kyoto University for their encouragement and discussions.

I also express my thanks to Professor *Masaaki Kawaguchi*, Professor *Atsuo Ono*, Associate Professor *Toru Nakai* and my colleagues at College of Liberal Arts of Kobe University for their guidance and encouragement in various ways.

Finally, I would like to express my gratitude to my wife *Atsuko* for her constant support and encouragement. My appreciation goes to my family for their encouragement for this work.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Integrated Communication Systems

The rapid progress in digital switching and transmission techniques offers a new opportunity to integrated communication systems which support a variety of information based services in the same network. Integrated communication systems accommodate all kinds of existing services such as telephone, facsimile and data transfer. They are also conceived to provide various future services such as video, broadcast communication and so on.

By the sharing of transmission and switching facilities, integration of a wide variety of applications will provide high system availability in a flexible and cost-effective manner. Compared to the alternative, which would be a set of parallel special-purpose networks such as conventional telephone networks and data communication networks, the concept of integrated communication systems has a number of advantages, particularly economy in planning, development, installation and operation. Special-purpose networks may require the installation of independent switching facilities for each service and hence a number of subscriber lines depending upon the services used. The integrated communication system, on the contrary, demands identified switching facilities for all the services and only one actual line per subscriber.

The fundamental problem of integration arises from the differences among various kinds of services both in traffic property and in transmission demand. That has resulted in a requirement for new transmission and switching technology of integrated networks. There are alternative approaches for the integrated communication system. As will be discussed in detail in Section 1.3, the most powerful means of integration is based on packet switching. In the packet switching network, all informations of applications are divided into small fractions constructed by packet format and each fraction is transmitted on the network.

This research thus focuses the attention on the performance evaluation of the effects of new transmission and switching techniques in integrated communication systems, where several performance measures and traffic models are introduced to evaluate individual characteristics of various services.

1.2 Nature of Problems in Integrated Communication Systems

1.2.1 Traffic Characteristics of Various Kinds of Services

Integrated communication systems are conceived to various kinds of existing and future communication services. Table 1.1 shows the representative potential services and their required transmission speeds. These service applications have different types of traffic characteristics and thus specified transmission requirements to maintain the quality and reliability of the services.

Voice

Conversational voice (=speech) communication will be as important in integrated communication systems as in the current telephone network.

Conversational voice communication needs to be transmitted in order and within some bound (typically 0.2 sec). The variability of the transmission delay may cause intermittent gaps in the reconstruction of speech at the receiver's end. Conversational voice, however, has high tolerance against transmission errors and loss of information. In voice conversation, the listener is not conscious of the loss and/or error of information up to 1 percent, and even if 10 percent of information is lost, he understands the message. Conversation thus does not require any loss/error detection and retransmission mechanism in the network.

High quality sound also requires fast and constant delivery just as conversational voice communication. However, it requires much more transmission quality than conversational voice. In such applications, packet retransmission is meaningless since it spends a lot of time, and then overhead bits for error protection mechanism should be attached to the packet at the end terminal.

Table 1.1 Representative potential services for integrated communication systems.

Type of Information	Service	Transmission Rate
Voice	Telephone Hi-Fi Sound	16 ~ 64 kbit/s 0.1 ~ 1 Mbit/s
Data	Text Large Data File Newspaper	1.2 ~ 64 kbit/s 0.1 ~ 1 Mbit/s 1 ~ 20 Mbit/s
Still Image	Facsimile Videotex Color Picture	0.05 ~ 1 Mbit/s 0.1 ~ 1Mbit/s 1 ~ 10 Mbit/s
Motion Image	Conventional TV High-Definition TV	50 ~ 100 Mbit/s 150 ~ Mbit/s

Data

Compared with conversational voice, data applications have much tolerance against transmission delay. They allow packets to be delayed variedly between end users, and the delay as much as a few seconds arises no problem. Data applications, on the other hand, require high reliability in packet delivery. For data communication, transmission error and packet loss must be detected, and damaged packet including discarded one must be retransmitted until it is successfully transmitted.

Still Image

Concerning transmission delay, still image has the same demands as data communications. As for transmission error and packet loss, the requirement depends on the applications of services. For example, facsimile and videotex services allow a small damage of information, but the service such as CAD images and computer graphics requires strict accuracy, and then error/loss detection and retransmission mechanisms are indispensable. We have to note that the future variable rate coding method using a new compression technique (see, for example, in [KISH89]) will make it possible to recover at least 1% loss without retransmitting a packet.

Motion Image

Motion image (=video) is the future applications for the broadband services,and has strict demands both for transmission delay and transmission damage. In the future network, the probability of transmission error will become very small because of the recent rapid advances in quality of transmission line, and the compression technique for image data will recover the small percentage of packet loss. In this sense, motion image has the same demands as conversational voice communication although the volume of information is very much larger than that of voice.

Classification of Services

The service applications described above are roughly categorized into two classes; *realtime* applications and *non-realtime* applications. Realtime applications require a fast packet delivery. As an alternative, they allow of transmission error and loss. Their quality should be supported by user's level (more than layer 4 as defined by the ISO Open System Interconnection Reference Model OSI/RM), if necessary. Non-realtime applications prefer accurate transmission to fast delivery. For such applications, network layer (layer 3) operates error/loss detection and packet retransmissions.

1.2.2 Traffic Patterns of Services

The services shown in Table 1.1 have also their own traffic patterns. In this section, we discusse traffic models which utilize the traffic patterns of the individual services.

Voice

Conversational voice (speech) is originally analog and is digitized by a coder at an end terminal. The most common coding technique today is the pulse code modulation (PCM), which is recommended to produce an 8-bit code every 125 micro seconds, resulting in an information transfer rate of 64 kbit/s [CCITT85a]. Other coding algolithms have been proposed for compressing to 32 kbit/s [CCITT85b] and below [RABI78], but with some degradation of the speech quality.

High quality sound is one of the future services, but high quality speech coding algolithms have been already proposed. The existing standard of CCITT [CCITT85c] involves sampling of 16kHz and 32kHz with converting each samples into 14 bit code. After including parity bits for error protection, the information transfer rates required are 192 and 384 kbit/s, respectively.

In packetized voice systems, samples of digital speech are accumulated in

packetizer for a period. If the stored samples have activity, they are packetized. Otherwise, they are discarded. Then, the voice traffic at a user is modeled as having alternating talkspurts and silences, with generation of voice packets at a constant rate during talkspurts and no packet generation during silence periods. Although the statistics of talkspurts and silences in speech depend on coding and packetizing mechanisms, it is usually modeled as a two-state Markov process [BRAD65], [MAXE82], [GRUB82b]. Being considered the situation that a lot of conversations are held under using the same transmission facility, the aggregated generation of voice packets becomes a Poisson process.

Data

Data traffic can be categorized into two basic types, interactive and bulk data. Interactive data such as remote time sharing service is bursty in nature. Bulk data such as file transfer has the nature of bulk packet generation for higher layer (higher than layer 4 of OSI/RM model), but it is not bulky for the layer which deals with network operation (lower than layer 3). In data communication, end-to-end flow control is employed at the layer, since strict error control and recovery procedure are required. For example, a simple window flow control protocol transmits a packet after confirming the successful transmission of a before-sending packet. Thus, in the viewpoint of network modeling, the generation of both types of data packets is modeled as a Poisson process.

Still Image

The volume of still image is much larger than that of typical data. For example, using straightforward coding, a 320 $\times$ 320 low resolution picture with 256 colors becomes about 0.5Mbit/picture, and 1000 $\times$ 1000 picture requires 1 Mbit/picture for a bilevel picture, 8 Mbit/picture for 256 gray levels, and 24 Mbit/picture for 256 levels each of red, green and blue. On the contrary, usual packet size is around the order of kbits. Therefore, one still image produces a lot of packets at once. The traffic patterns for still image are classified as three patterns depending on both a transmission demand and transmission speed of physical line. If the application has tolerance against transmission and if the transmission speed of the line is fast compared with the processing speed of the terminal, the pattern is modeled just as conversational voice packets. If the transmission speed is not so fast under the same transmission demand, the pattern is modeled as bulk arrivals obeying a Poisson process. The third pattern is used when an application requires a strict accuracy. In this case, the image is treated just as data packets, thus the packet generation is assumed to obey a Poisson process.

Motion Image

Motion image requires much higher transmission rate than other services. Straightforward digitization of a video signal based on the National Television System Committee (NTSC) quality requires about 86-115 Mbit/s; using differential pulse code modulation (DPCM) this rate can be reduced to 44.7 Mbit/s [O'NEL86]. A full bandwidth HDTV signal occupies 1.2 Gbit/s; compression technique may reduce this to 100-300 Mbit/s; and future advances of coding technology will reduce this much less.

Since a motion image requires the above transmission rate during whole service period, straightforward coding generates a packet every short constant time period. It should be noted that capacity of the switching facilities which can support such high bit rate applications is very large (may be much more than several Gbit/s), and then the packet generation does not look bulky. Indeed the traffic pattern of packets using compression technique depends on its compressing mechanism, but it is out of the scoop of this research.

1.2.3 Performance Measures

As mentioned in 1.2.1, service applications in integrated communication systems have different types of transmission requirements. To examine the performance for these requirements, representative performance measures are listed below.

Throughput / Channel Utilization

Since one of the major motivations for integration is to make efficient use of transmission channels, *throughput* is a basic performance measure. It is defined in this thesis as the average numbers of correctly transmitted packets per unit time. When the throughput is normalized by one packet transmission time on a channel (if all packets are assumed to have the same packet length), it is called *channel utilization*. Channel utilization is a convenient measure to estimate the efficiency of channel use, since the maximum value of channel utilization for each channel is equal to 1.

Transmission Delay

Average transmission delay is an important performance measure. In many integrated networks, especially in wide area networks, packets are transmitted via several switching nodes. Two types of transmission delay arise in such networks. When we pay attention to an end-to-end level transmission, an end-to-end transmission delay, which is called a *traverse time* in Chapter 2, is defined as a time interval from the time a packet arrives at a source terminal to the time at which it is received at a destination terminal. When we pay attention to a behavior of a certain switching node, *transmission delay of the switching node* is defined as the time interval from the time a packet arrives at the switch to the time at which it is transmitted on an outgoing channel.

For traffic of realtime applications, the *distribution of the transmission delay* is also an important measure. For instance, in conversational voice communication, the percentage of packet arrivals at destination within a predetermined transmission delay is more important factor than average transmission delay.

Packet Loss Probability

For realtime applications, packet loss probability is a key factor to evaluate the quality of services, since the retransmission of discarded packets is not employed.

Discarded packets of non-realtime applications must be retransmitted until they are successfully transmitted. Then packet loss causes a large end-to-end transmission delay for end users because the retransmission takes a lot of time. The retransmission may also cause the packet congestion of network, since the offered traffic load of the network is increased by retransmission. Therefore, packet loss probability is an important factor for both types of applications.

1.3 Approach to the Integrated System Design and Its Performance

1.3.1 Switching Alternative

We have two basic switching techniques for conventional communication networks; circuit switching and packet switching. Circuit switching is used as a telephone network and has the most rigid channel structure, while packet switching, which is popular in data communication networks, has the most flexibility. Recently, much of the research interest has tended to focus on techniques capable of supporting variable bandwidth services in the same switching fabric. The fabric technology which seems to be securing the most interest is based on packet switching.

In the rest of this section, we will briefly review the previous approaches to the integrated packet switching network design.

1.3.2 Integration onto Store-and-Forward Packet Switching Networks

One way to construct an integrated network is to use an existing packet switching network which was built for specific purpose. Store-and-forward type packet switching is the most popular switching technique in the conventional data communication network. In the network, each transmission channel operates at a speed of 64-784 kbit/s, then the interest is focused on the integration of relatively low speed data and conversational voice.

The first approach was started by ARPA project in 1973. Some experiences were conducted over ARPANET [CASN78]; in 1974, realtime packet voice communication was demonstrated over ARPANET between ISI (California) and Lincoln Laboratory (Massachusetts) and in 1976 online voice conferencing capabilities were demonstrated among ISI, Lincoln Laboratory, CHI (Santa Barbara) and SRI (Stanford). From theses experiments, we have found the necessity of investigating the behaviors of voice packets in more detail.

In the previous works which dealt with the performance analysis of the integrated store-and-forward networks, the network is modeled as a simple queueing network. A single switching node is modeled as a M/M/1 queueing system in [COVI77] and [FORG77] and as a nonpreemptive priority M/G/1 system with two distinct input queues for voice and data in [MOWA80]. In these works, both data and voice packets were simply assumed to have Poisson arrivals and an identical distribution of packet size.

1.3.3 Integration onto Local Area Networks

Local Area Networks (LANs) have developed as small area networks such as those in a same building, factory, campus and so on. With the requirement of the international standardization, IEEE presented the standards for LANs which deal with the physical and data link layers of OSI model. Following the implementations of those standards, though they have flexibility to be revised and reaffirmed, many commercial LANs which operate a moderate speed (3-32 Mbit/s) transmission channel, have been prepared for data applications. Now our interest has focused on the use of standard data communication LAN for integration of realtime applications, especially integration of packetized voice.

Integration onto CSMA/CD LANs

Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method is one of the access methods of IEEE 802 standards [IEEE85a]. CSMA/CD network is one of the most popular and prominent local area networks, and has high reliability with simple transmission protocol. However, it causes much variance in transmission delay which is unavoidable due to the undeterministic nature of contention and collision backoff [TOBA80]. Gonsalves [GONS82] and DeTreville [DeTR84] experiment the feasibility of

realtime voice transmission on CSMA/CD LAN using 3 Mbit/s experimental Ethernet and the simulation study, respectively, but the native problem under CSMA/CD that the variance of transmission delay grows much more rapidly than its mean is still unresolved.

From its importance, some works have been done to estimate the capability of realtime voice transmission and to get higher performance on a revised version of CSMA/CD method. The typical protocols of them are summarized as follows.

1) assign higher retransmission rate to voice packets, when collision occurs [JOHN81],[NUTT82],[MATS88].
2) incorporate non-contention type algorithm such as token passing protocol into a back-off algorithm [RIOS85].
3) add extra preamble bits to voice packets to give priority to them [MAXE82],[TOBA82].
4) give access right to voice group and data group alternately, and voice users transmit packets using a proper protocol for realtime delivery during their access right (i.e., use a hybrid protocol) [SEN87].
5) use a framed hybrid protocol; divide a channel into fixed length frames each of which consists of a voice subframe and a data subframe. Voice users transmit their packets during voice subframe [CHLA85], [MEDI85], [OKAD84].

Integration onto Token Passing Ring LANs

Token passing ring network is the other popular LAN for IEEE 802 standards [IEEE85b]. Since the media access method of token passing ring is based on non-contention algolithm, it has a large capability to support realtime applications. Despite of its importance little has been known about the performance of a token passing ring network when it is subjected to a voice load in additions to a data load. Due to the realtime constraints of interactive voice applications, it is not apparent to us if the token passing ring network offers a variable alternatives to other methods of interconnecting users.

Integration onto High-Speed LANs

Another interesting field of integrated LANs is to construct a high speed integrated LAN for future network. The Fiber Distributed Data Interface (FDDI) is a proposed American National Standard for 100 Mbit/s token ring using an optical fiber medium [ROSS86]. FDDI, which was originally proposed as a packet switching network, adds a circuit switching capability, expanding the field of application of FDDI to include voice and video applications.

Orwell ring is another proposed protocol which is based on the slotted ring principle with slots released at the destination node. Simulation experiment in [FALC85] showed that a 140 Mbit/s Orwell ring can carry over 170 Mbit/s of speech whilst keeping these delays under 200 micro sec.

1.3.4 Integration onto High-Speed Packet Switching Networks

The current wide area networks which use the store-and-forward packet switching technique can support the bit rate at most 1 Mbit/s. Thus, new high-speed transfer techniques which employ the high speed transmission of channel bit rates greater than 100 Mbit/s have been required to offer broadband service applications such as video applications. CCITT has stated the first baseline documents for Broadband Integrated Service Network (B-ISDN). The Asynchronous Transfer Mode (ATM) is considered the ground on which B-ISDN is to be built. The fundamental concept of ATM was proposed by Hasegawa et al., in [HASE64] and [HASE66], although the situation considered there (e.g., transmission speed, service application) was far from today's situations. In ATM, all information to be transmitted is divided into fixed-sized packets called 'cells', which are identified and switched by means of

a label in the header.

Since many of ATM functions are not yet defined, the ATM approach requires many new problems to be solved. One of the important subjects is to establish a switch architecture for B-ISDN in a high speed environment. The switch is needed to handle a lot of channels each operating at a speed of more than 100 Mbit/s. In order to realize such high speed switching, a new switching technique based on hardware switching has been required. The proposed switching architectures are roughly categolized into three groups; banyan type [HUI87],[TURN86], bus-matrix type [NOJI87], [YEH87] and ring type [TAKE87]. These are based on the single channel allocation policy. Pattavina [PATT88] introduced a multichannel allocation scheme in which a set of channels between two switching nodes is viewed as a single virtual channel by routing entities. He presented a switch architecture implementing the scheme in Batcher-banyan switch. The multichannel allocation scheme has an advantage of solving the congestion problem in the switch, moreover, the scheme makes an efficient use of transmission channels. Thus, the scheme is conceived to be promising for future very high speed B-ISDN network.

1.4 Overview of the dissertation

The major objective of this dissertation is to investigate high performance integrated communication systems and to evaluate their performances. Chapter 2 contributes the integration onto store-and-forward packet switching network. Chapter 3, 4 and 5 offer the integration onto local area network. Chapter 6 contributes the future switching technique towards B-ISDN. In what follows, the contents of each chapter are summarized.

Chapter 2 investigates a voice and data integration on store-and-forward packet switching system which would be modeled as a type of tandem queueing system with two classes of packets. A voice packet which is expected to suffer from longer delay than predetermined duration is lost at each switching node, and there is no preferential treatment between voice and data packet. But, the length of voice packet is different from that of data packet because of their characteristics. The steady state probabilities of each node are analyzed exactly when possible and are approximated otherwise. The performance measures obtained include traverse time (end-to-end transmission delay in the network) and loss probability for voice packet due to nodal buffer capacity. With the proposed methods, it has become possible to evaluate the characteristic quantities of the system considered.

Chapter 3 proposes and analytically evaluates a hybrid VC-CSMA/CD protocol in integrated local area networks. A fixed length frame consists of a virtual circuit (VC) subframe and random access (RA) subframe, which are separated with movable boundary. In a VC subframe, connection oriented stations such as voice stations, transmit packets in a virtual circuit fashion, which ensures a collision free and realtime packet delivery. In an RA subframe, random access stations such as data stations, access the channel according to the asynchronous non-persistent CSMA/CD protocol. Two effective and practical VC techniques are proposed and analyzed using

queueing models. The analytical method for the throughput and delay of RA traffic and the loss probability of VC calls is derived.

Chapter 4 evaluates the performance of a token passing ring network through extensive simulations to explore the feasibility of this class of networks as a medium for combined voice and data transmission. Both data and voice users are modeled in simulations. The data users present bursty traffic. Voice traffic is modeled as having alternating talkspurts and silences, with generation of packets at a constant rate during talkspurts and no packet generation during silence periods. Performance measures obtained include: the distribution of the transmission delay, loss probability of voice packets, the number of voice users allowed on a network to satisfy the realtime requirements of speech.

Chapter 5 considers a transportation type bulk-arrival bulk-service queueing system with composite bulk service discipline as a fundamental research on future integrated communication networks. In this system, packets arrive in groups and packets in a group have an identical attribute. From the head of the queue, packets with the same attribute are served within a finite bulk size at the same time. This service discipline can be applicable to the integrated LANs with token-based access methods such as token passing ring, token passing bus and FDDI.

For such bulk service queueing system, the average queue length of packets in the system is approximately analyzed. In our approach, each class of packets with the same attribute is assumed to form their own queue. The approximation method derives queue length of each queue just before the service initiation epoch, where probability generating function approach and embedded Markov chain method are utilized.

Chapter 6 proposes a new high-speed packet switch architecture based on a Channel Group Virtual Circuit (CG-VC) scheme for integrated communication networks. We consider the case in which there are several parallel channels between two switching nodes. The CG-VC scheme allows a packet to select an available channel in such a way as to avoid congested channels using a simple, hardware-based, self routing mechanism. As a result, the architecture provides efficient channel use and low switching delay. We also evaluate the performance of the proposed switch by approximation analysis and simulations. Performance measures obtained include channel utilization (=throughput) and average switching delay in a switch.

Chapter 7 gives some concluding remarks and discussion for the future research.

The results discussed in Chapter 2 are mainly taken from [OHTS84], Chapter 3 from [OHTS89], Chapter 4 from [SUDA85], Chapter 5 from [OHTS85], and Chapter 6 from [OHTS90, OHTS91a].

CHAPTER 2

Performance Analysis for Traverse Time in an Integrated Store-and-Forward Packet Switching Network

2.1 Introduction

With the development of computer networks, it has been desired to establish integrated communication services in which various kinds of information flows are transmitted on the same communication channel. Especially, integration of voice and data is of great interest. One of the most powerful means for integration is the packet switching used in standard data communication networks. Not only in data communication but also in voice communication, traffic becomes bursty on conversation although each call is not bursty. Packet switching is also the most cost-efficient technique for transmitting bursty traffic. Some experiments have been conducted over the ARPANET [CASN78], [COVI77], [GITM77]. However, in the above circumstances, different types of characteristics, which bring with them many problems intractable, prevent the idea from realizing.

Major differences between voice communication and data communication are as follows [CANS78], [COHE81], [FORG75], [FORG77], [GITM77], [GITM78], [GRUB81], [GRUB82a].

a) Transmission delay

Data can be transmitted variedly between terminals and be delayed for a few seconds. Voice needs to be transmitted in order and within fixed delay (typically 200 ms).

b) Error detection

Concerning data packets, transmission error must be detected and retransmission must be performed. For voice communication, error detection can be omitted.

c) Loss of information

In data communication it is not permitted. In voice communication, the listener is not conscious of the loss of information up to 1 percent, and even if 10 percent of information is lost, he understands the message.

d) Volume of information

Generally, volume of data information is greater than that of voice information.

In short, for the voice packets, transmission delay should be less than a certain level but they have much more tolerance against transmission error than data packets. Data packets should be transmitted with low error rate but have very high tolerance against transmission delay.

End-to-end delay of voice traffic is defined as the sum of three major components, packetization delay at the sending terminal, network queueing delay and playout time delay at the receiving terminal. Packetization delay is determined by both vocoding rate and packet length. Playout delay depends on reassembly strategy. Only network delay depends on the degree of network congestion. In [COVI77], [FORG77], [MINO79] and [MOWA80], network delay is analyzed as a simple queueing model. A single integrated node is modeled as an M/M/1 queueing system [COVI77], [FORG77] and as a nonpreemptive priority M/G/1 system with two distinct input queues for voice and data [MOWA80].

This chapter considers network delay in more detail. The network considered consists of many switching nodes interconnected by high speed transmission lines, where both voice and data are accommodated by the store-and-forward packet-switched concept. The special features of our model are as follows[OHTS84].

1. A voice packet which is expected to suffer from longer delay than predetermined duration is lost at each node in the network.
2. There is no preferential treatment between voice packet and data packet.

3. The length of a voice packet is different from that of a data packet because of their characteristics.
4. Each switching node has a single server.

Priority systems like [MOWA80] have higher performance for voice packet than our model, but they need new switching hardware and/or software. What we are concerned with here is the case that voice and data traffic are integrated under the standard packet communication network.

The network considered would be modeled as a type of tandem queueing system where a voice packet is lost when the buffer is full. In a global sense, this model is a kind of store-and-forward network model with finite nodal storage capacity. Lam [LAM76] and Schweitzer et al. [SCHW76] consider a multi class packet model. From queueing theoretical point of view, each queue has a single server with exponential service time under Kleinrock's independence assumption [KLEI64], that is, message lengths can be treated to be independent of interarrival times in a communication network. But in our model, service time does not obey exponential distribution because of the difference of packet length between voice and data. We assume the service time of voice packets and data packets to be exponentially distributed with their own parameter μ_1 and μ_2, respectively.

The key factors for the performance of this system are traverse time (end-to-end delay in the network) and loss probability for voice packet due to nodal buffer capacity.

The steady state probabilities of each node are analyzed exactly if possible and approximated otherwise. Approximations are validated in comparison with the results of simulation. And the numerical results show that buffer limit can suppress the traverse time, but the shorter buffer limit is cut out, the more voice packets are lost. With the proposed method, it has become possible to obtain the optimal buffer limit and the bound for utilization.

2.2 Model

The network model considered here consists of many switching nodes interconnected by high speed transmission lines, where data and voice are integrated. There is no preferential treatment between voice and data packets, and service discipline at each node is FCFS (first come first served). Each node has finite buffer, and when the buffer of the next node is full, an arriving packet is lost if it carries voice or retransmitted after a random time if it carries data. Since rapidity of transmission is important, it will be justified to dispose the voice packet which is expected to arrive at a destination too late when the buffer is full. That will avoid congestion and let the other packets arrive earlier, and only the loss should be controlled within a certain level.

In our model we consider the case that conversations have been held through several switching nodes. All voice packets generated from the speaker are transmitted on the same route in the network to maintain the order. When a voice call occurs, one virtual circuit between a couple of transmitter and receiver will be set up through N queueing nodes as shown in Fig. 2.1. We assume that all conversations have been held between node 1 and node N, and that there is no other virtual circuits in the sequential network.

Voice packet and data packet may have different packet length. Generally speaking, voice packet is much shorter than data packet. If the voice packet is longer, not only network delay but also packetization delay becomes longer. On the contrary, if data packet is short, overhead (i.e. header, tailer) is large in comparison with real information part, consequently network congestion increases. All actions for transmission are treated as the services, and a service time for each packet is considered to be in proportional to the length of packet. We assume the service time of voice packets and data packets to be exponentially distributed with their own parameters μ_1 and μ_2, respectively.

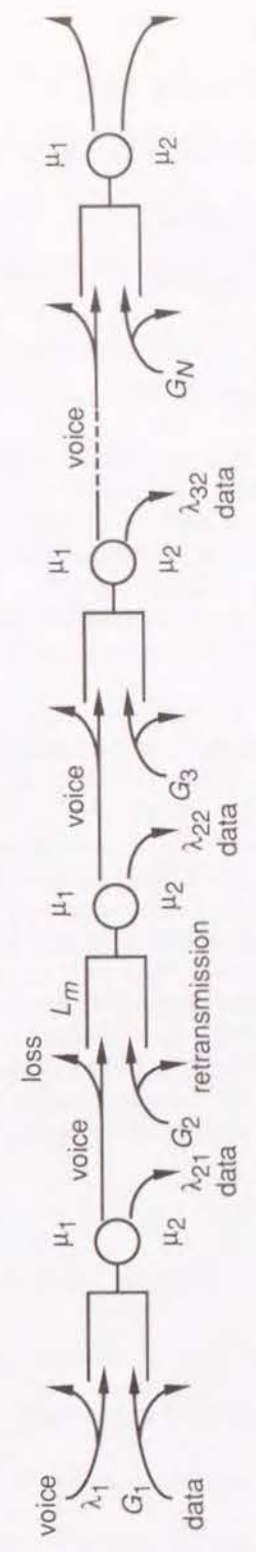


Fig. 2.1 Network model.

We define following notations as shown in Fig. 2.1 .

λ_1 : arrival rate of voice packet at node 1,

λ_{2i} : arrival rate of new data packet at node i,

G_i : arrival rate of new and retransmitted data packet at node i,

$1/\mu_1$: mean service time of voice packet,

$1/\mu_2$: mean service time of data packet,

L_m : maximal allowable queue length including a packet in service at each node (buffer limit + 1).

Let it be supposed that a data packet falls out of this sequential network after completion of a certain service. To be more precise, put the case that data packet transmitted from node i to node $i+1$ would be included in the arrival of data from the outside to node $i+1$.

An interarrival process of data packet is assumed to be Poisson process, and that of voice packets in conversation is as well. The reason is that each voice message is packetized and that only a part of talkspurt is transmitted. If it is assumed that the retransmission interval of data packet is long compared to interarrival interval, the effect of the retransmission attempts is to magnify the Poisson arrival stream. In the following analysis the arrival process of data packet including retransmission one is assumed to be a Poisson process.

2.3 Analysis

The traverse time is defined as the time interval necessary for voice packet to be sent from a source node to a destination node. At first, the stochastic behavior at the first node is analyzed by using Markov chain analysis, where arriving packets forms a Poisson process. Next the one at the second node is analyzed. And then the analysis of the stochastic behaviors of the other nodes is considered. The equilibrium equations for the steady state probabilities at these nodes can be easily described such as nodes one and two, but their dependency on the previous nodes causes the large size of states in arithmetical progression, which makes it impossible to calculate the state probabilities by a present computer. In more detail, the number of the different states increases in proportion to $(L_m{}^2+L_m+1)^N$, where N denotes the number of the switching nodes. Then we propose two approximation methods to obtain the steady state probabilities of each node. One is that the arrival process of the tagged node is simply approximated as a Poisson process, and the other is that the arrival process of the previous node is approximated as a Poisson process. Therefore, if the steady state probabilities of node i are obtained, then that of node $i+1$ can be approximated to obtain the mean delay of voice packet at node $i+1$ and so on.

2.3.1 Stochastic Behavior of Node 1

Fig. 2.2 shows the model of node 1. The state probability, $P(i, j, k)$, denotes the probability that a packet of class k ($k=1$, or 2) is in service, and that i packets of class 1 (=voice) and j packets of class 2 (=data) are waiting for service. $P(i, j, k)$ can be obtained by using M/H_2/1 model. Our model is very similar to M/H_2/1 model shown in Fig. 2.3, where a packet arrives according to exponential distribution with parameter λ_1+G_1 , and as for service he chooses service stage 1 with probability $\lambda_1/(\lambda_1+G_1)$ or chooses stage 2 with probability

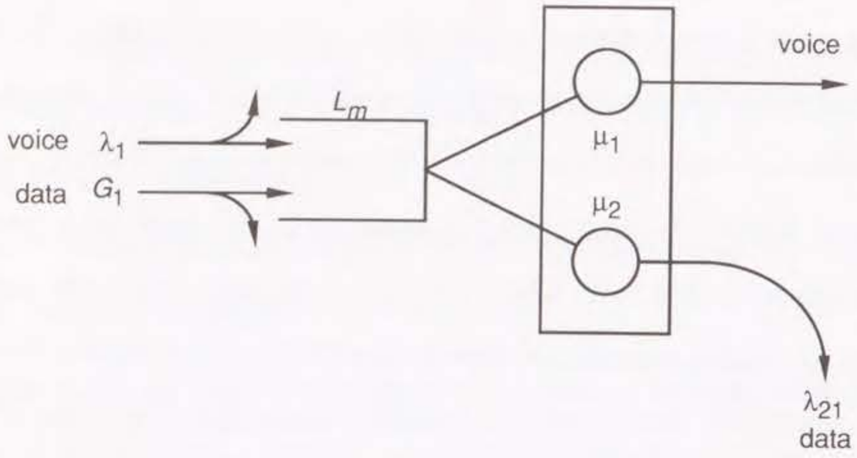


Fig. 2.2 Queueing model of node 1.

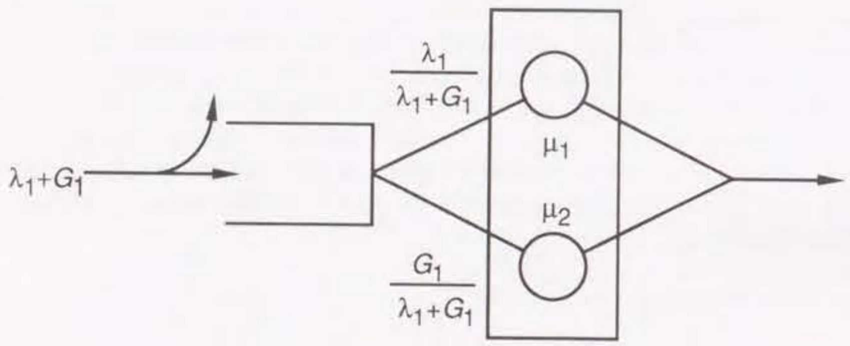


Fig. 2.3 M/H$_2$/1 model.

$G_1/(\lambda_1+G_1)$. Then the packet will be given an exponentially distributed service in the i-th stage with mean $1/\mu_i$ ($i=1, 2$). The difference between the two models is that class 1 packets and class 2 packets are distinguished in our model but not in M/H$_2$/1 model. If we denote the state probability $\pi(i, k)$ as the probability that i packets are in system with class k ($k=1, 2$) packet in service, then $P(i, j, k)$'s become

$$P(i,j,k) = \binom{i+j}{j} A^j (1-A)^j \, \pi(i+j+1, k) \qquad (k=1,2) \tag{2.1}$$

$$P(0,0,0) = \pi(0,0) \tag{2.2}$$

where

$$A = \lambda_1/(\lambda_1+G_1)\ . \tag{2.3}$$

Equation (2.1) indicates that when N ($N=i+j$) packets are waiting for service, the probability that there are i packets of class 1 in the queue obeys binomial distribution $B(N, \lambda_1/(\lambda_1+G_1))$, which is obvious according to the memoryless property of exponential distribution. State probability $\pi(i, k)$ is explicitly obtained by M/H$_2$/1 formula [WHIT75], as a function of unknown variable G_1.

The loss probability of both data and voice packets at node 1, $P_{loss}(1)$, is,

$$P_{loss}(1) = \sum_{k=1}^{2} \sum_{(i,j)\in X} P(i,j,k) \tag{2.4}$$

where

$$X = \{(i,j) \mid i+j=L_m-1\}\ . \tag{2.5}$$

And the relation between G_1 and λ_{21} is described by

$$G_1 = \lambda_{21}/(1-P_{loss}(1)) \tag{2.6}$$

State probabilities are obtained by solving the above simultaneous equations (2.1)-(2.6). But unfavorably these equations don't form linear equations, which make it quite cumbersome to calculate. Our approach to solve them uses iteration method and Markov chain method in a following sense. We pay attention to the character that the bottleneck equation is eq. (2.6), in other words, eqs. (2.1)-(2.5) are linearly related. Then, at first, we ignore eq. (2.6) and numerically analyze state probabilities from eqs. (2.1)-(2.3) by using Markov method where λ_{21} is substituted for G_1. Next, calculate $P_{loss}(1)$, and then compute G_1 from eq. (2.6). Next, substitute the value, G_1, got from eq. (2.6) for G_1 of eqs. (2.1)-(2.3), calculate again and get the new value of G_1. Then compare the new value, $G_1(new)$, with previous one, $G_1(old)$, and iterate above calculation until the difference between them is negligible. About this criterion we will discuss in the following section.

The Laplace transform of delay time distribution of voice packet at node 1, $L_1(s)$ is

$$L_1(s) = \Bigg(P(0,0,0)\left(\frac{1}{\mu_1+s}\right) + \sum_{(i,j)\in Y} P(i,j,1)\left(\frac{\mu_1}{\mu_1+s}\right)^{i+2}\left(\frac{\mu_2}{\mu_2+s}\right)^{j}$$

$$+ \sum_{(i,j)\in Y} P(i,j,2)\left(\frac{\mu_1}{\mu_1+s}\right)^{i+1}\left(\frac{\mu_2}{\mu_2+s}\right)^{j+1}\Bigg) / (1-P_{loss}(1)) \tag{2.7}$$

where,

$$Y = \{(i,j) \mid i+j<L_m\}\ . \tag{2.8}$$

Mean delay time for voice packet and i-th moment of its distribution $E[D_1{}^i]$ is given by

$$E(D_1^i) = (-1)^i \frac{d^i}{ds^i} L_1(s)\ \Big|_{s=0} \tag{2.9}$$

2.3.2 Stochastic Behavior of Node 2

Now, we consider a two node model shown as in Fig.2.4. The steady state behavior of node 2 is independent of the other nodes but node 1. Here, the behavior of node 2 is analyzed. Following notation is used,

$P(i, c_1, j, k, c_2)$: steady state probability that i packets are in system with a class c_1 packet in service at node 1, and j packets of class 1 and k packets of class 2 are waiting for service while class c_2 packet is in service at node2. Note that when $i=0$, c_1 is defined as 0.

The equilibrium equations are described as follows.

$$A(0,0,0,0,0)\,P(0,0,0,0,0) = \mu_2 P(1,2,0,0,0) + \mu_1 P(0,0,0,0,1) + \mu 2P(0,0,0,0,2) \tag{2.10}$$

$$A(1,c_1,0,0,0)\,P(1,c_1,0,0,0) = B(0,c_1)\,P(0,0,0,0,0) + \mu_1 P(1,c_1,0,0,1) + \mu_2 C(c_1)\,P(2,2,0,0,0) + \mu_2 P(1,c_1,0,0,2) \tag{2.11}$$

$$A(i,c_1,0,0,0)\,P(i,c_1,0,0,0) = B(i-1,c_1)\,P(i-1,c_1,0,0,0) + \mu_1 P(i,c_1,0,0,1) + \mu_2 C(c_1)\,P(i+1,2,0,0,0) + \mu_2 P(i,c_1,0,0,2) \qquad (\text{if } 1< i \leqq L_m) \tag{2.12}$$

$$A(1,c_1,j,k,1)P(1,c_1,j,k,1) = B(0,c_1)P(0,0,j,k,1) + \mu_1 C(c_1)P(2,1,j-1,k,1) + \mu_2 C(c_1)P(2,2,j,k,1) + \mu_1 D(1)P(1,c1,j+1,k,1) + \mu_2 D(1)P(1,c_1,j+1,k,2) + G_2 P(1,c_1,j,k-1,1) \qquad (\text{if } j+k<Lm-1) \tag{2.13}$$

$$A(\mathrm{i},c_1,j,k,1)P(i,c_1,\mathrm{j},k,1) = B(i-1,c_1)P(i-1,c_1,j,k,1) + \mu_1 C(c_1)P(i+1,1,j-1,k,1) + \mu_2 C(c_1)P(i+1,2,j,k,1) + \mu_1 D(1)P(i,c_1,j+1,k,1) + \mu_2 D(1)P(i,c_1,j+1,k,2) + G_2 P(i,c_1,j,k-1,1) \qquad (\text{if } i=0 \text{ or } 1< i \leqq L_m, \text{ and } j+k<L_m-1) \tag{2.14}$$

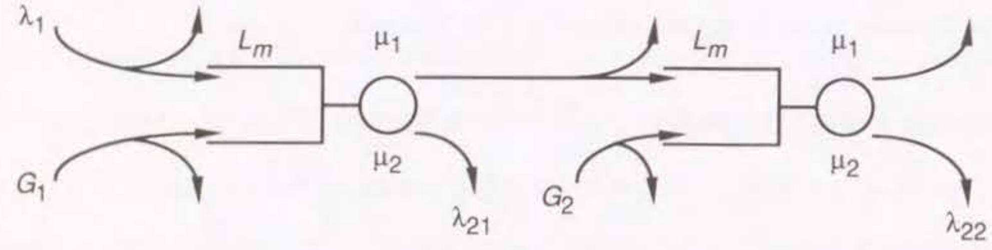


Fig. 2.4 Two node model.

$$A(1,c_1,j,k,2)P(1,c_1,j,k,2) = B(0,c_1)P(0,0,j,k,2) + \mu_1 C(c_1)P(2,1,j-1,k,2)$$
$$+ \mu_2 C(c_1)P(2,2,j,k,2) + \mu_1 D(2)P(1,c_1,j,k+1,1) + \mu_2 D(2)P(1,c_1,j,k+1,2)$$
$$+ G_2 P(1,c_1,\mathrm{j},k-1,2) \qquad (\text{if } j+k<L_m-1) \tag{2.15}$$

$$A(i,c1,j,k,2)P(i,c_1,j,k,2) = B(i-1,c_1)P(i-1,c_1,j,k,2) + \mu_1 C(c_1)P(i+1,1,j-1,k,2)$$
$$+ \mu_2 C(c_1)P(i+1,2,j,k,2) + \mu_1 D(2)P(i,c_1,j,k+1,1) + \mu_2 D(2)P(i,c_1,j,k+1,2)$$
$$+ G_2 P(i,c_1,j,k-1,2) \qquad (\text{if } i=0 \text{ or } 1<i\leqq L_m, \text{ and } \mathrm{j}+k<L_m-1) \tag{2.16}$$

$$A(1,c_1,j,k,c_2)\mathrm{P}(1,c_1,j,k,c_2) = B(0,c_1)P(0,0,j,k,c_2) + \mu_1 C(c_1)P(2,1,j-1,k,c_2)$$
$$+ \mu_1 C(c_1)P(2,1,j,k,c_2) + \mu_2 C(c_1)P(2,2,j,k,c_2) + G_2 P(1,c_1,j,k-1,c_2)$$
$$(\text{if } j+k=L_m-1) \tag{2.17}$$

where

$$P(i,c1,j,k,c2) = 0 \qquad (\text{if } i<0,\ j<0,\ k<0,\ i>L_m \text{ or } j+k \geqq Lm), \tag{2.18}$$

$$A(i,c_1,j,k,c_2) =$$
$$\begin{cases} \lambda_1 + G_1 + G_2 + \mu_{c_1} + \mu_{c_2} & (\text{if } 0\leqq i\leqq L_m \text{ and } 0\leqq j+k<L_m-1) \\ \lambda_1 + G_1 + \mu_{c_1} + \mu_{c_2} & (\text{if } 0\leqq\ <L_m \text{ and } j+k=L_m-1) \\ G_2 + \mu_{c_1} + \mu_{c_2} & (\text{if } i=L_m \text{ and } 0\leqq j+k<L_m-1) \\ \mu_{c_1} + \mu_{c_2} & (\text{if } i=L_m \text{ and } i+j=L_m-1), \end{cases} \tag{2.19}$$

$$B(i,c_1) = \begin{cases} \lambda_1 + G_1 & (\text{if } 1\leqq i\leqq L_m) \\ \lambda_1 & (\text{if } i=0 \text{ and } c_1=1) \\ G_1 & (\text{if } i=0 \text{ and } c_1=2), \end{cases} \tag{2.20}$$

$$C(c_1) = \begin{cases} \lambda_1/(\lambda_1+G_1) & (\text{if } c_1=1) \\ G_1/(\lambda_1+G_1) & (\text{if } c_1=2) \\ 1 & (\text{if } c_1=0), \end{cases} \tag{2.21}$$

$$D(c_2) = \begin{cases} (j+1)/(j+k+1) & (\text{if } c_2=1) \\ (k+1)/(j+k+1) & (\text{if } c_2=1) \end{cases} \tag{2.22}$$

$$\mu_{c_1} = \begin{cases} \mu_1 & (\text{if } c_1=1) \\ \mu_2 & (\text{if } c_1=2) \\ 0 & (\text{if } c_1=0) \end{cases} \tag{2.23}$$

μ_{c_2} is defined in the same way. Normalizing condition is as follows,

$$\sum_{i,c_1,j,k,c_2} P(i,c_1,j,k,c_2) = 1 \tag{2.24}$$

The probability that the data packet is lost at node 2, $P_{loss}(D_2)$ is given by

$$P_{loss}(D_2) = \sum_{i,c_1,c_2} \sum_{j,k \in X} P(i,c_1,j,k,c_2) \tag{2.25}$$

The relation between G_2 and λ_{22} is,

$$G_2 = \lambda_{21}/(1-P_{loss}(D_2)) \tag{2.26}$$

The calculation method to obtain steady state probability is similar to that of sec. 2.3.1. The value of G_1 is already known if state probabilities of node 1 have been calculated. Then we calculate eqs. (2.10)-(2.25) substituting λ_{22} for G_2 and iterate using similar algolithms to that of sec. 2.3.1.

The probability that the voice packet is lost at node 2, $P_{loss}(2)$, is described by

$$P_{loss}(2) = \frac{\sum_{i,c_2} \sum_{(i,j) \in X} P(i,1,j,k,c_2)}{\sum_{i,c_2} \sum_{(i,j) \in Z} P(i,1,j,k,c_2)} \tag{2.27}$$

where,

$$Z = \{(i,j) \mid i+j \leqq L_m - 1\}. \tag{2.28}$$

The Laplace transform of delay distribution of voice packet at node 2, $L_2(s)$, is given by

$$\begin{aligned} L_2(s) = \Bigg(& \sum_{i=1}^{L_m} P(i,1,0,0,0)\left(\frac{1}{\mu_1+s}\right) \\ & + \sum_{i=1}^{L_m} \sum_{(i,j) \in Y} P(i,1,j,k,1)\left(\frac{\mu_1}{\mu_1+s}\right)^{j+1}\left(\frac{\mu_2}{\mu_2+s}\right)^{k} \\ & + \sum_{i=1}^{L_m} \sum_{(i,j) \in Y} P(i,1,j,k,2)\left(\frac{\mu_1}{\mu_1+s}\right)^{j+1}\left(\frac{\mu_2}{\mu_2+s}\right)^{k+1} \Bigg) \\ & / \left(\sum_{i=1}^{L_m} \sum_{(i,j) \in Y} \sum_{c_2=1}^{2} P(i,1,j,k,c_2) + \sum_{i=1}^{L_m} P(i,1,0,0,0) \right) \end{aligned} \tag{2.29}$$

The steady state probability of node 2 is defined as $P_2(j, k, c_2)$, which is similar to $P(i, j, k)$ of section 2.3.1.

<u>Approximation method of node 2 (APPRO-1)</u>

In the previous discussion, exact method for the steady state probability of node 2 is proposed. But from a practical point of view, many variables which are used to express the state probability cause long computation time or large amount of working area at the time of calculation. Herein, the simple approximation method is proposed where the arrival process of voice packet at node 2 is assumed to confirm Poisson process with parameter λ' which is obtained by

$$\lambda' = \sum_{i=1}^{L_m} \pi(i,1) \tag{2.30}$$

Then node 2 is approximately analyzed similarly to section 2.3.1. We call this approximation APPRO-1.

2.3.3 Stochastic Behavior of Node n

As mentioned before, it is difficult to obtain the stochastic behavior of node n ($n>2$) because of its dependency on the previous nodes. If the arrival process is given, however, the steady state probabilities of the node can be described. The arrival process of voice packet at node n ($n=3, \ldots. N$) is equal to the departure process of that from node $n-1$. We propose two approximation methods for steady state probabilities of node n, where the departure process of voice packet of node $n-1$, or node $n-2$ is approximated as a Poisson process. The former approximation is equal to the APPRO-1 described in section 2.3.2. In the latter approximation method the arrival distribution of voice packet at node i is considered to have more accuracy than that of APPRO-1. In this approximation model the arrival process of voice packet at node $n-1$ confirms a Poisson process with parameter $\mu_1 P_a(n)$, where $P_a(n)$ denotes the probability that a voice packet is in service at node $n-2$. The equilibrium equation is obtained using (2.1)-(2.24), where λ_1, λ_{21}, λ_{22}, G_1 and G_2 are exchanged to $\mu_1 P_a(n)$, λ_{2n-1}, λ_{2n}, G_{n-1} and G_n respectively. $P_a(n)$ is obtained if the steady state probabilities of node $n-2$ are analyzed, and described by

$$P_a(n) = \sum_{i=1}^{L_m} \sum_{(i,j,) \in Y} \sum_{c_1=1}^{2} P_{n-2}(i, c_1, j, k, 1) \qquad (2.31)$$

where $P_{n-2}(i, c_1, j, k, 1)$ is the steady state probability of nodes $n-2$ and $n-3$.

Note that in this method G_{n-1} is not equal to that given by analyzing node $n-1$ because of approximation error. Then we have to converge this value at first, and next converge G_n using iteration method.

The characteristics of node i can be analyzed similarly to section 2.3.2. We call this approximation APPRO-2.

2.3.4 Traverse Time

Traverse time is defined as the time experienced by a voice packet in the network. The mean traverse time $E[T]$ is approximated as

$$E[T] = \sum_{k=1}^{N} E[D_k] \qquad (2.32)$$

where $E[D_k]$ is the mean delay at node k.

We assume that the delay time distribution of each node is independent of the other nodes. Then the Laplace transform of traverse time distribution, $L_T(s)$, is given by

$$L_T(s) = \prod_{k=1}^{N} L_k(s) \quad . \qquad (2.33)$$

The i-th moment, $E[D_T{}^i]$ is given by

$$E[T] = \sum_{k=1}^{N} (-1)^i \frac{d^i}{ds^i} L_T(s) \mid_{s=0} \quad . \qquad (2.34)$$

The probability that a voice packet is lost in the network is denoted by $P_{loss}(T)$, which is described as

$$P_{loss}(T) = 1 - \prod_{i=1}^{N} (1 - P_{loss}(i)) \quad . \qquad (2.35)$$

2.4 Numerical Results

In the following examples, the criterion to stop iteration is set up by

$$1-\frac{G_i(old)}{G_i(new)}<10^{-5}.$$

In the analytic model arrival process of data packets including retransmitted ones at each node is assumed to be a Poisson process, but in the simulation, it is assumed more practically as follows. The data packet which was not accepted to enter the buffer retries to transmit after a random retransmission interval, and continues this trial until it is accepted. The retransmission interval obeys the exponential distribution with mean $1/\mu_r$. If μr is very little compared to λ_{2i}, both analytic and simulation model are almost the same. Following numerical calculations are performed by $\lambda_{2i} = \lambda_2$ $(i=1,\ldots,N)$ and $\mu_r = \mu_2/10$ unless otherwise provided.

Our approximation results are compared with simulation data. Then, at first, the simulation results are compared with exact ones, for the reason of testing validity of the simulation experiment. Table 2.1 shows the comparison of the exact results and the simulation ones for node 1. The simulation results are well verified by the theoretical results.

We consider, next, the packet lengths of data and voice packets. In general, the length of a voice packet is assumed to be of the order of 300-700 bits, and that of data packet in the integrated network is assumed to be of the order of 1000-2000 bits. The service rate is inversely proportional to a packet length. Table 2.2 shows that as the ratio μ_1/μ_2 increases, the loss probability of voice packet also increases. In the following numerical examples the ratio of service rate is fixed to 4.

Fig. 2.5 shows the mean delay of voice packet at node 1 ($E[D_1]$) as a factor of arrival rate of voice packet λ_1. In low traffic range, all cases have almost the

Table 2.1 Comparison between exact analysis and simulation.

			$E[D_i]$	$E[D_i^2]$	$E[D_i^3]$	$P_{loss}(i)$
$\lambda_1=.8$ $\lambda_2=.8$ $\mu_1=8.$ $\mu_2=2.$ $L_m=7$	node 1	Analysis	.3387	.3589	.6693	.2558E−2
		Simulation	.3384	.3592	.6749	.3516E−2
		C.I. (95%)	±.0040	±.0094	±.0329	±.3601E−3
	node 2	Analysis	.3449	.3643	.6754	.4016E−2
		Simulation	.3449	.3643	.6737	.4068E−2
		C.I. (95%)	±.0033	±.0086	±.0283	±.2827E−3
$\lambda_1=1.$ $\lambda_2=.5$ $\mu_1=8.$ $\mu_2=2.$ $L_m=5$	node 1	Analysis	.3175	.3078	.5286	.1351E−1
		Simulation	.3205	.3104	.5313	.1376E−1
		C.I. (95%)	±.0240	±.0263	±.0526	±.7903E−3
	node 2	Analysis	.3216	.3104	.5303	.1442E−1
		Simulation	.3205	.3080	.5245	.1436E−1
		C.I. (95%)	±.0245	±.0266	±.0519	±.7941E−3

C.I. : Confidence Interval

Table 2.2 Comparison of the ratio $\mu_1:\mu_2$.

$(\lambda_1=1.0,\ \mu_1=8.0)$

$\mu_1:\mu_2$	L_m		$E[D_1]$	C.V.	$P_{loss}(1)$
4 : 1 ($\lambda_2=1.$) ($\mu_2=2.$)	5	Analysis	.3204	1.431	.1390E−1
		Simulation	.3209	1.434	.1391E−1
	10	Analysis	.3480	1.482	.4869E−3
		Simulation	.3496	1.490	.4987E−3
8 : 1 ($\lambda_2=.5$) ($\mu_2=1.$)	5	Analysis	.4494	1.794	.3252E−1
		Simulation	.4521	1.792	.3272E−1
	10	Analysis	.5327	1.797	.3232E−3
		Simulation	.5305	1.812	.3216E−3

same delay. But as λ_1 increases the system becomes saturated in the case Lm=20. On the contrary, in the cases $L_m=3$ and $L_m=5$, the delay increase almost proportionally with the arrival rate. It must be considered that the case of $L_m=20$ is not suitable for this network. On the other hand, the shorter Lm (= buffer limit) is, the more the loss probability of voice packet becomes (see Fig. 2.6). In the case of $L_m=20$, the loss probability is almost equal to 0, but in the case of $L_m=3$, it becomes very large if the arrival rate λ_1 exceeds 0.5 . Since this probability must be suppressed within 0.05, shorter buffer limit is not suitable.

As seen in Fig. 2.7, the relation of the arrival rate to the coefficient of variation of delay distribution (C.V.) is very complex. This indicates that many factors (arrival rate, buffer limit, the ratio of voice packet against all packets, etc.) affect the delay distribution.

Table 2.3 shows two examples of the stochastic behavior of each node, where proposed approximations are compared with simulation results. Parameters of those cases are as follows.

Case (a) $\lambda_1=\lambda_2=0.8$, $\mu_1=8.0$, $\mu_2=2.0$, $L_m=5$.

Case (b) $\lambda_1=\lambda_2=0.5$, $\mu_1=8.0$, $\mu_2=2.0$, $L_m=5$.

With respect to the mean delay, $E[D_i]$, analytical results obtained by APPRO-2 method is well verified by simulation results. In general, $E[D_i]$ increases with the number of node. It seems that the arrival process affects the delay strongly.

Table 2.4 shows the simulation results in which the influence of buffer limit is ignored (i.e. infinite buffer). The C.V. of the departure distribution is larger than 1. On the contrary, that of data packet is smaller than 1 . Note that in the case of exponential distribution the C.V. becomes equal to 1. Herein, it is said that the larger C.V. of arrival distribution becomes, the more delay increases.

$E[D_i]$ of APPRO-1 decreases by all means as the node number i increases, which is owing to the diminution of arrival rate.

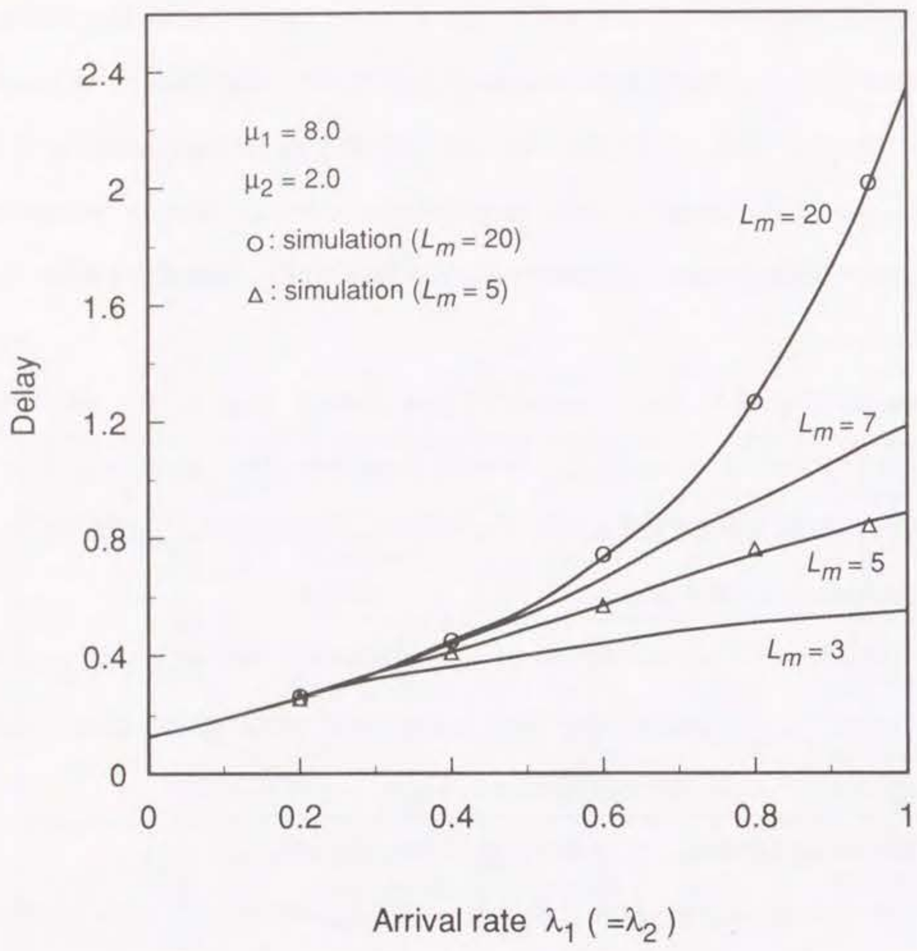


Fig. 2.5 Average delay of voice packet at node 1.

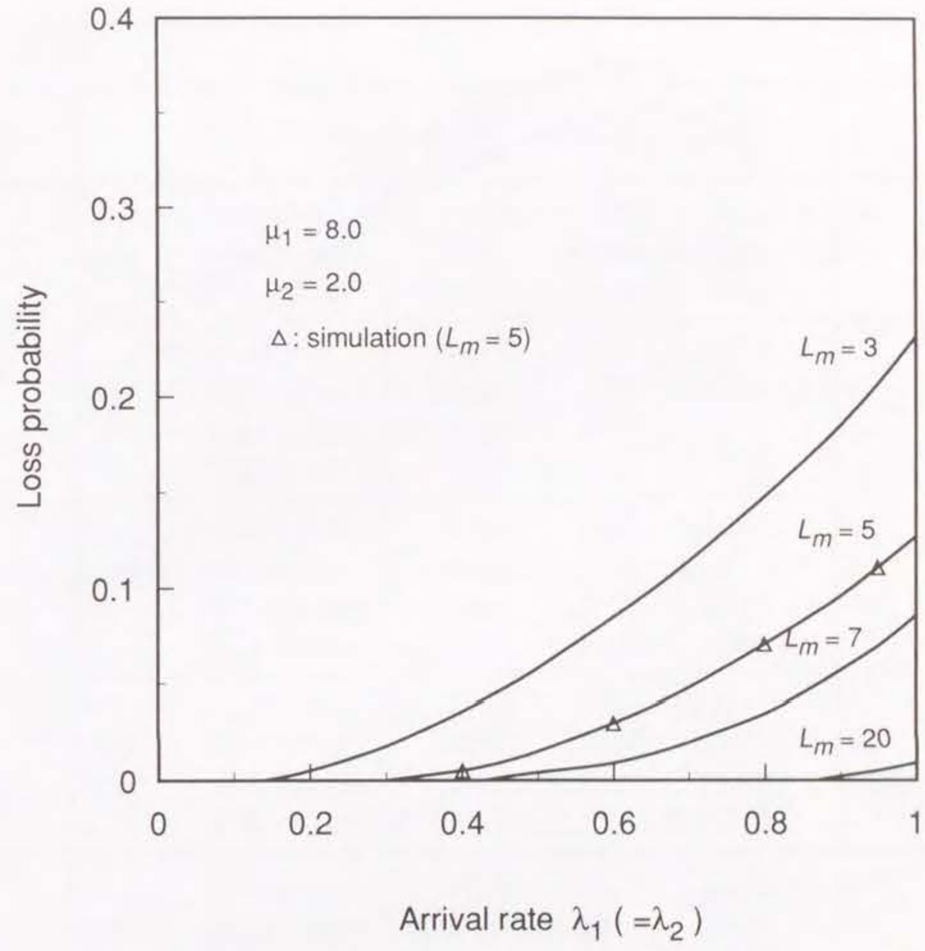


Fig. 2.6 Loss probability of voice packet at node 1.

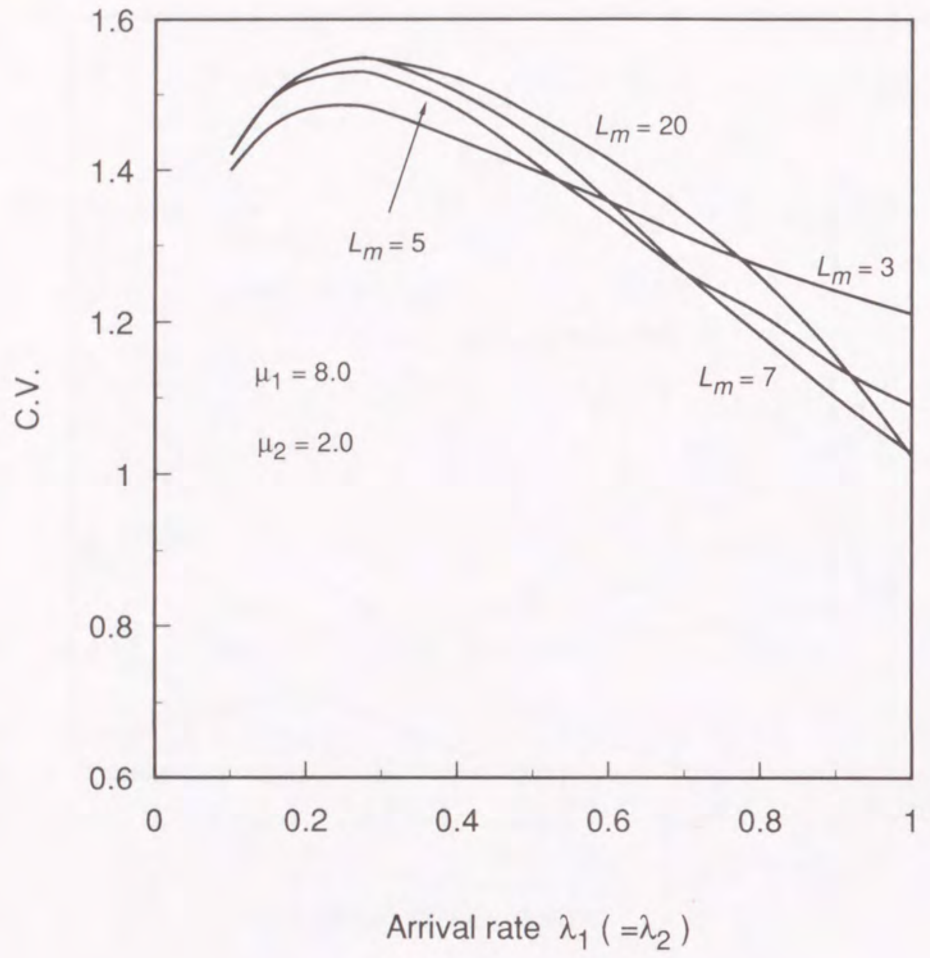


Fig. 2.7 Coefficient of variation of voice delay distribution at node 1.

Table 2.3 Stochastic behaviors at each node.

			$E[D_i]$	$E[D_i^2]$	$E[D_i^3]$	C.V.	$P_{loss}(i)$
$\lambda_1=.8$, $\lambda_2=.8$, $\mu_1=8.$, $\mu_2=2.$, $L_m=5$	node 1	Analysis	.4570	.5750	1.1474	1.324	.3177E−1
		Simulation	.4583	.5820	1.1762	1.331	.3217E−1
		C.I. (95%)	±.0358	±.0519	±.1199		±.0105E−1
	node 2	Analysis	.4605	.5767	1.1474	1.311	.3334E−1
		APPRO-1	.4563	.5751	1.1496	1.327	.3075E−1
		Simulation	.4639	.5847	1.1677	1.310	.3397E−1
		C.I. (95%)	±.0361	±.0514	±.1172		±.0094E−1
	node 3	APPRO-2	.4599	.5767	1.1496	1.314	.3235E−1
		APPRO-1	.4559	.5751	1.1517	1.330	.2980E−1
		Simulation	.4598	.5732	1.1393	1.309	.3503E−1
		C.I. (95%)	±.0367	±.0528	±.1206		±.0112E−1
	node 4	APPRO-2	.4593	.5768	1.1517	1.316	.3136E−1
		APPRO-1	.4552	.5752	1.1537	1.332	.2982E−1
		Simulation	.4663	.5840	1.1544	1.298	.3528E−1
		C.I. (95%)	±.0365	±.0531	±.1218		±.0098E−1
$\lambda_1=.5$, $\lambda_2=.5$, $\mu_1=8.$, $\mu_2=2.$, $L_m=5$	node 1	Analysis	.3061	.3025	.5363	1.493	.5254E−2
		Simulation	.3041	.3024	.5180	1.487	.5134E−2
		C.I. (95%)	±.0228	±.0244	±.0487		±.0329E−2
	node 2	Analysis	.3087	.3043	.5378	1.481	.5591E−2
		APPRO-1	.3060	.3024	.5362	1.493	.5219E−2
		Simulation	.3081	.3038	.5338	1.484	.5144E−2
		C.I. (95%)	±.0230	±.0244	±.0482		±.0383E−2
	node 3	APPRO-2	.3087	.3042	.5378	1.481	.5554E−2
		APPRO-1	.3059	.3023	.5362	1.493	.5184E−2
		Simulation	.3108	.3038	.5400	1.473	.5721E−2
		C.I. (95%)	±.0231	±.0253	±.0472		±.0351E−2
	node 4	APPRO-2	.3086	.3042	.5378	1.481	.5517E−2
		APPRO-1	.3059	.3023	.5362	1.494	.5362E−2
		Simulation	.3127	.3079	.5438	1.466	.6184E−2
		C.I. (95%)	±.0233	±.0253	±.0532		±.0435E−2

C.I. : Confidence Interval

Table 2.4 Stochastic behaviors at each node (infinite buffer).

			Departure			Delay
			m_1	m_2	C.V.	$E[D_1]$
$\lambda_1=.8$	voice	node 1	1.2502	3.3961	1.0831	.5522
$\lambda_2=.8$		node 2	1.2502	3.6281	1.1495	.5620
$\mu_1=8.$		node 3	1.2502	3.8383	1.2065	.5737
$\mu_2=2.$		node 4	1.2503	4.0205	1.2538	.5771
$L_m=7$	data	node 1	1.2470	3.0612	.9842	.9289
		node 2	1.2476	3.0633	.6737	.9309
		node 3	1.2507	3.0699	.9811	.9348
		node 4	1.2462	3.0437	.9796	.9341

m_i : i-th moment, C.V. : coefficient of variation.

Concerning the loss probability, both approximation results seem to have tendency to underestimate the simulation ones, but the results obtained by APPRO-2 method indicate the better values compared with that by APPRO-1.

Table 2.5 shows the characteristic quantities for tandem queues consisting of 4 nodes, where, C.V., Utilz and $P(T>3.5)$ indicate the coefficient of variation for the traverse times, mean utilization factor and the probability that the traverse time exceeds 3.5, respectively. APPRO-2 result seems to come to near by exact value.

Note that the mean traverse time is not a summation of mean delay at each node. For example, simulation results show that in case (a) of Table 2.3, the summation of mean delay is 1.842 (= .4583 + .4639 + .4598 + .4663), but the corresponding mean traverse time is 1.868 shown in Table 2.5.

If we set the limit of the traverse time to 3.5, the rate of successful voice packet is obtained as $1-P_{loss}(T)-P(T>3.5)$.

Summarily, many results show that the APPRO-2 method provides high approximation accuracy, especially on the delay of a node, and that APPRO-1 method can be used as a simple approximation method to know the behavior for network delay.

Table 2.5 Stochastic behaviors for end-to-end transmission .

	L_m		$E[T]$	C.V.	*Utilz.*	$P_{loss}(T)$	$P(T>3.5)$
$\lambda_1=1.$ $\lambda_2=.5$ $\mu_1=8.$ $\mu_2=2.$	7	APPRO-1	1.357	.7296	.3739	.1413E−1	
		APPRO-2	1.377	.7207	.3739	.1558E−1	
		Simulation	1.396	.7145	.3734	.1624E−1	.4278E−1
	5	APPRO-1	1.279	.7166	.3708	.5285E−1	
		APPRO-2	1.293	.7090	.3078	.5634E−1	
		Simulation	1.303	.7037	.3681	.5887E−1	.3120E−1
	3	APPRO-1	1.079	.7033	.3587	.1985	
		APPRO-2	1.083	.6994	.3587	.2024	
		Simulation	1.104	.7001	.3493	.2058	.1532E−1
$\lambda_1=.5$ $\lambda_2=.5$ $\mu_1=8.$ $\mu_2=2.$	7	APPRO-1	1.262	.7591	.3124	.3416E−2	
		APPRO-2	1.272	.7540	.3124	.3663E−2	
		Simulation	1.282	.7520	.3127	.4090E−2	.3578E−1
	5	APPRO-1	1.224	.7466	.3117	.2065E−1	
		APPRO-2	1.232	.7419	.3117	.2175E−1	
		Simulation	1.238	.7385	.3108	.2200E−1	.2938E−1
	3	APPRO-1	1.077	.7242	.3075	.1249	
		APPRO-2	1.080	.7419	.3076	.1272	
		Simulation	1.092	.7219	.3022	.1298	.1583E−1
$\lambda_1=.8$ $\lambda_2=.8$ $\mu_1=8.$ $\mu_2=2.$	7	APPRO-1	2.025	.6747	.4974	.4053E−1	
		APPRO-2	2.047	.6673	.4976	.4382E−1	
		Simulation	2.060	.6576	.4946	.4747E−1	.1406
	5	APPRO-1	1.825	.6640	.4925	.1158	
		APPRO-2	1.839	.6580	.4929	.1232	
		Simulation	1.868	.6555	.4681	.1296	.1025
	3	APPRO-1	1.424	.6659	.4776	.3323	
		APPRO-2	1.429	.6628	.4784	.3414	
		Simulation	1.462	.6656	.4552	.3467	.4315E−1

2.5 Conclusion

In this chapter, we have suggested a tool for the performance evaluation for integrated store-and-forward packet switching systems. An integrated communication network for packetized voice and data has been reduced to a series queueing model in which a voice packet is lost when the buffer of the next stage is full. The steady state probabilities of the first two nodes have been analyzed exactly, but those of the other nodes can not be exactly calculated because the number of different states increases in arithmetical progression. Then they have been approximated by two approximation methods and validated through comparison with the results of simulation. With the proposed method, it has become possible to evaluate the characteristic quantities of the system considered.

Voice communication requires much less delay than data traffic, but it can torelate more transmission error. While data packets should have low error rate but have very high torelance against transmission delay. When a voice packet suffers from longer delay than predetermined duration, this packet should be lost. Therefore, traverse time and the loss probability due to longer delay for the voice packet are the key factors for the performance of the system.

If the buffer limit of each node is cut short in order to decrease the traverse time, the loss probability of voice packet increases. On the other hand, the longer the packet queue length is allowed to reduce loss probability, the longer the traverse time becomes. Then the utilization of this network must be reduced lower than a normal data network so as to suppress the traverse time and the loss probability to a lower level.

With the method proposed here, the interactions of voice packet transmission and data packet transmission have been revealed to fairly high extent, although many problems such as relation between transmission delay of voice packet and the articulation of the decoded voice signal at the receiving end remain as a future work.

CHAPTER 3

Performance Analysis of a Hybrid VC-CSMA/CD Protocol in Integrated Local Area Bus Networks

3.1 Introduction

With the requirement of the international standardization, IEEE presented the standards for local area networks (LANs) which deal with the physical and data link layers as defined by the ISO Open System Interconnection Reference Model. Following the implementations of those standards, though they have flexibility to be revised and reaffirmed, many commercial LANs have been prepared for data applications. Now our interest has focused on the use of standard data communication LAN for integration of realtime packetized voice which is traditionally transmitted in virtual circuit (VC) fashion.

Carrier Sense Multiple Access with Collision Detection (CSMA/CD) access method is one of the access methods of IEEE 802 standard [IEEE85a]. CSMA/CD network is one of the most popular and prominent local area networks, and has high reliability in nature. That is, the media access control operation provides simpler asynchronous access than other protocols such as token passing ring, slotted ring, and quickly recovers from channel failure. Moreover, it does not depend upon the number of stations nor the locations of them. Therefore, and for the economical reasons as well, the protocol is suitable for small area networks such as those in a same building or a little campus. However, it causes much variance in transmission delay which is unavoidable due to the undeterministic nature of contention and collision backoff [TOBA80]. Gonsalves [GONS82] and DeTreville [DeTR84] experiment the feasibility of real-time voice transmission on a CSMA/CD LAN using 3 Mbit/s experimental Ethernet and the simulation study, respectively, but the native problem under CSMA/CD that the variance of transmission delay grows much more rapidly than its mean is still unresolved.

From its importance, some works have been done to estimate the capability of real-time transmission and to get higher performance on a revised version of CSMA/CD method. The typical protocols of them are summarized as follows.

1) assign higher retransmission rate to voice packets, when collision occurs [JOHN81],[NUTT82],[MATS88].
2) incorporate non-contention type algorithm such as token passing protocol into a back-off algorithm [RIOS85].
3) add extra preamble bits to voice packets to give priority to them [MAXE82],[TOBA82].
4) give access right to voice group and data group alternately, and voice users transmit a packet using a proper protocol for real time delivery during their access right (i.e., use a hybrid protocol) [SEN87].
5) use a framed hybrid protocol; divide a channel into fixed length frames each of which consists of a voice subframe and a data data subframe. Voice users transmit their packets during voice subframe [CHLA85], [MEDI85], [OKAD84].

Among them, the framed hybrid protocol seems to be the most desirable and hopeful protocols for real-time packet delivery on bus networks. Especially, Clamtac [CLAM85] proposed the transmission protocol which ensure the constant transmission delay of real-time applications without any interference of bursty data traffic during a connection. It means that, if the packet transmission interval of a voice station is equal to the sampling interval of voice signal at a packetizer, voice packets are delivered as if they are in dedicated circuits. Consequently, no specific play-out strategy is required at a

destination station. A station plays out voice packets one by one upon their arrivals. Another feature of his protocol is that it makes an efficient channel use, since the boundary of the subframes is movable depending on the volumes of real-time packets. However, his protocol has some problems which are caused by a hand-shake mechanism of real-time packet. One is that the protocol requires the $2D$ (D is a maximum end-to-end delay) additional overhead to each voice packet, which will prevent an efficient channel utilization on high speed channel. Second is that the hand-shake mechanism is rather complicated for high-speed transmission; especially, in the case that a station does not work accurately for some accident, it is difficult to control and recover the hand shaking procedure.

As for the performance analysis for framed hybrid protocols, little has been done in spite of its importance. In [MEDI85], data stations are assumed to access the channel according to slotted, non-persistent CSMA (not CSMA/CD) in a fixed length data subframe for analytical convenience.

In this chapter, we investigate a hybrid VC-CSMA/CD protocol in integrated local area networks[OHTS89]. A fixed length frame consists of a virtual circuit (VC) subframe and random access (RA) subframe, which are separated with movable boundary. In a VC subframe, connection oriented stations such as voice stations, transmit packets in a virtual circuit fashion, which ensures a collision free and real-time packet delivery. In an RA subframe, random access station such as data stations, access the channel according to the asynchronous non-persistent CSMA/CD protocol. Two effective and practical VC techniques are proposed. These protocols have advantages of 1) simple operation mechanisms, 2) making much efficient channel use (overhead between packets is reduced to D), and 3) having much reliability for damage of a station.

We also analytically evaluate the proposed protocols using a more practical queueing models with movable boundary. The analytical method for the throughput and delay of RA traffic and the loss probability of VC calls is derived.

In the next section, we consider the transmission protocol where two practical VC techniques are proposed. We derive the analytical method for the loss probability of VC call and throughput of data traffic and in section 3.3 and 3.4. In section 3.5, we consider the performance for both types of traffic using analytical and simulation results.

3.2 Hybrid VC-CSMA/CD Protocol

The network considered in this paper is assumed to consists of two kinds of stations. One treats with connection oriented traffics and transmits a packet with virtual circuit fashion, and the other treats with connectionless (datagram) traffics and transmits a packet with random access fashion. In this chapter, we call the former a VC station and the later a RA station.

3.2.1 Frame Structure

Consider a time axis described as Fig.3.1. A single channel is divided into successive frames of fixed length, each of which consists of a virtual circuit subframe (VF), a random access subframe (DF) and channel control control periods. A VC station transmits a packet on VF, and a RA station transmits a packet on DF. In the remainder of this chapter, we call the time duration of a frame, frame time, for convenience.

3.2.2 Channel Division Mechanism

A transmission mode controller (TMC) has a timer and transmits a "voice subframe starting" signal (VFS) every frame time whether the channel is idle or not. VFS has enough preambles in order to prevent the VFS from colliding and get the channel strictly. That is, the preambles allow the transmitting source, if any, to detect a collision and terminate transmission before TMC begins transmitting a useful informations. When VF is terminated, TMC sends a "voice subframe ending" signal (VFE), and the transmission mode is changed to RA packet transmission mode (i.e., RF). One of the practical method for this kind of channel division technique is proposed in [OKAD84].

3.2.3 Voice Packet Transmission Mechanism

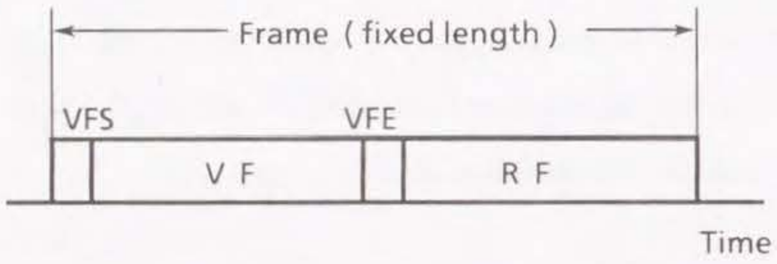


Fig. 3.1 Frame structure.

In a voice subframe, voice stations transmit voice packets in a virtual circuit (VC) fashion, which ensures a collision free and real-time packet delivery. We consider the following three practical techniques. In those techniques, the channel access controls are negotiated between voice sources. Voice sources set timing clock on the basis of VFS, if necessary, which means that the protocol is tolerant of timing errors of VFS.

Movable Boundary with TDM (M-TDM) Protocol

A VF is partitioned into minislots with each minislot having enough length to transmit a packet. The station which has been setting up a circuit is called a ready station. The ready station is ensured to transmit a packet in each VF until it releases the circuit. In this method, all ready stations use successive minislots in order, which provides effective channel utilization by controlling VF lengths dynamically. More precisely, the method creates and maintains the successive virtual circuits as follows.

A ready station remembers its transmission ordering number and has a transmission right at a corresponding minislot in each VF. For example, the ready station whose ordering number is i (say station i) can use the i-th minislot. When the station i terminates a call, it sends a "release circuit" signal using the tailer of its last packet. Each ready station knows the fact of station i by sensing the channel and moves up the ordering number by one at the end of VF if its current number is greater than i.

The station whose ordering number is the last (say last station) in VF knows the fact and sends "end of circuit (EC)" signal following to its packet at the end of the minislot and then an access period starts. The stations who want to create the circuits sense at least one VF and send access packets immediately after detecting EC. If only one station accesses the channel during a access period, it succeed in setting up a circuit and makes itself a last station in the next VF. The current last station decides whether it will still the last station or not in the next VF, by observing the access period. Note that if two or more stations access the channel during an access period, a collision occurs and, as a result, no one creates a circuit. However, this situation occurs very rarely, because the frame length is on the order of tens of milliseconds while the mean connection time is on the order of hundreds of seconds (see the discussion in section 3.5 in detail). The collided stations give up or retry accessing according to a proper algolithms. When TMC detects EC, it sends VFE after waiting for the end of an access period, and VF is terminated.

In this method, the length of VF changes dynamically depending on the number of ready stations. To ensure a certain measure of transmission for connectionless traffic the maximum allowable number of circuits is required. When a station finds that there are no more available circuits in the next VF by observing a current one, the station gives up accessing the channel at this subframe.

Movable Boundary with Asynchronous TDM (M-ATDM) Protocol

This protocol is the revised version of M-TDM protocol in which every ready station monopolizes a predetermined minislot of fixed length. On the contrary, M-ATDM protocol achieves asynchronous virtual circuits. That is, a ready station is allowed to transmit a packet of flexible length (of course maximum size is limited), and moreover all packets in VF are transmitted consecutively. More precisely, the method works as follows.

A ready station has a transmission ordering number as M-TDM protocol. Each ready station senses the channel and counts the number of transmitted packets in VF. Station i transmits a packet after detecting the completion of packet transmission of station $i-1$. If there is no information to be transmitted at this time, station i transmits short size packet which consists of header plus tailer, in order to maintain the transmission order.

3.2.4 RA Packet Transmission Mechanism

RA stations can use conventional non-persistent CSMA/CD controllers by adding a special adopter in the receiver. The adopter sends its controller busy tone during VF, hence, the controller regards a VF as if there is a long packet on the channel. It should be noted that 1-persistent CSMA/CD protocols recommended in IEEE std. 802.3 is not suitable for the hybrid protocol, because many data packet arrived during a VF may be waiting for transmission at the beginning of data subframe which result in colliding in use of 1-persistent CSMA/CD.

As TMC transmits VFS regardless of the channel activity on DF, the packet in transmission may be interfered by VFS at the end of DF. And what is worse, the following situation may happen : a packet is interfered by VFS before the destination receives it completely, though the source station terminates the transmission. To prevent this situation, the source station senses the channel after transmission and when it finds VFS within $2D$ it considers the transmission was interfered. The interfered packet is retransmitted afterwards.

3.3 Analysis for VC Traffic

In the hybrid VC-CSMA/CD protocol, the length of virtual circuit subframe (VF) does not depend on the status of RA traffic, and VC stations can transmits packet without any interference during its channel holding period. On the contrary, the length of random access subframe (RF) strictly depends on the length of VF, consequently, the performance of RA traffic is affected by the status of VC traffic. In the analysis, we then analyze the stochastic behaviors of VC stations at first.

Since VC packet are ensured to be transmitted with constant delay once the circuit is setup, then our interest is focused on evaluating the validity of the access method for virtual circuit. As its performance measure, we derive the probability of failures when it tries to set up the circuit.

As for RA stations, the length of VF is an important measure, since it affects the length of RF directly (note that the sum of VF length and RF length is constant.). In this section we also derive the probability distribution of VF length.

3.3.1 VC Traffic Model

A channel of which transmission capacity is C is divided into successive frames of length T_f along time axis. The propagation delay between stations is assumed to be constant of length D.

We assume infinite VC stations. A VC station, which generates a requirement of connection for VC, tries to set up the circuit at the end of VF. A time interval required to access for this trial is called access period, and is denoted as t_{ac}. The station which has failed to set up a circuit is called backoff station, and it retries to access the channel after a random time until it successfully sets up a circuit. The offered traffic load not only of new connection requirements but also of retrying ones is assumed to be a Poisson

stream with parameter G_v. The number of available virtual circuits is limited to N_v. And connection holding time obeys a exponential distribution with mean $1/\mu$.

At a transmission epoch, VC station transmits a packet of the constant length L_m, if the station has some information to be transmitted. Thus, the minislot length of M-TDM is the sum of the packet transmission time L_v/C and propagation delay D. In M-ATDM, VC station transmits a short packet of length H, if there is no information to be transmitted.

3.3.2 Analysis

Concerning the probability of fault for connection and the number of VCs, both proposed protocols have the equivalent stochastic quantity, for given traffic parameters. But the length of VF is different between two protocols. Thus,we at first analyze the common characteristics, then derive individual characteristics.

Let p_i be the steady state probability that i VC stations are under establishing the VC. In order to derive p_i, we denote the following notations.

- q_s : probability that only one station accesses at an access period (eventually the station success the set up of VC).
- P_{re} : probability that a certain ready station terminates its connection and releases the VC after at a VF.
- $\mu_v(i,j)$: probability that j out of i ready stations release the VCs in VF.

They then follow that

$$q_s = G_v \cdot T_f \cdot exp(-G_v \cdot T_f) \tag{3.1}$$

$$P_{re} = 1 - exp(-\mu \cdot T_f) \tag{3.2}$$

$$\mu_v(i,j) = \begin{cases} \binom{i}{j} P_{re}^{\ j}(1-P_{re})^{i-j} & (\text{if } i \geqq j \geqq 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.3}$$

Equilibrium equations of p_i's are described using eqs. (3.1)-(3.3) as,

$$(1-(1-q_s)\,\mu_v(i,0)-q_s\cdot\mu_v(i,1))\,p_i$$
$$=\sum_{j=1}^{N_v-i}[(1-q_s)\,\mu_v(i+j,j)+q_s\,\mu_v(i+j,j+1)]\,p_{i+j}+q_s\,\mu_v(i-1,0)\,p_{i-1}\;. \tag{3.4}$$

From eq. (3.4) and the normalizing condition (i.e., $\Sigma\, p_i=1$), we obtain the state probabilities p_i's.

When a voice station wants to access an access minislot, it gives up accessing the minislot due to being full of circuits with probability P_{full}, and it fails to set up a circuit due to collision with probability P_{coll}. P_{full} and P_{coll} are described by,

$$P_{full}=p_{N_v}\mu_v(N_v,0) \tag{3.5}$$

$$P_{coll}=1-exp(-G_v\cdot T_f)\quad . \tag{3.6}$$

Then the probability that a voice station fails to set up the channel when it accesses an access minislot, P_{fault}, is given by

$$P_{fault}=1-(1-P_{full})(1-P_{coll})\quad . \tag{3.7}$$

If we denote the throughput S_v as the average number of VCs per unit time, S_v is given by

$$S_v=G_v(1-P_{fault})\quad . \tag{3.8}$$

[Distribution of VF Length in a M-TDM protocol]

Let $x(i)$ be the length of VF when the number of VCs in a VF is i. Since a VF consists of i minislots and one access period, it is given by,

$$x(i)=(L_v/C+D)\cdot i+t_{ac} \tag{3.9}$$

The VF length is distributed discretely depending on the number of VCs. If we denote $P_{lv}(X)$ as the probability that the length of VF is equal to X, it is obtained by

$$P_{lv}(X)=\begin{cases} p_i & (\text{ if } X=x(i)\,) \\ 0 & (\text{ otherwise}) \end{cases} \tag{3.10}$$

[Distribution of VF length in a M-ATDM protocol]

We assume that each ready station generates a packet with probability α at each transmission epoch. Then a ready station transmits an actual packet of length L_v with probability α, and transmits a short packet of length H with probability $1-\alpha$. If we denote $P(i,j)$ as the probability that j out of i ready stations transmit actual packets, $P(i,j)$ is given by

$$P(i,j)=\begin{cases} \binom{i}{j}\alpha^j(1-\alpha)^{i-j} & (\text{ if } i\geqq j\,) \\ 0 & (\text{ otherwise }) \end{cases} \tag{3.11}$$

Let $X(i,j)$ be the length of VF when j out of i ready stations transmit actual packers, $X(i,j)$ becomes

$$x(i,j)=(L_v/C+D\,)\cdot j+(H/C+D\,)\,(i-j)+t_{ac} \tag{3.12}$$

The distribution of VF length is described by

$$P_{lv}(X)=\sum_{(i,j)\in \mathrm{ij}_X} p_i\cdot P(i,j) \tag{3.13}$$

where

$$\mathrm{ij}_X=\{\,(i,j)\mid x(i,j)=X\} \tag{3.14}$$

3.4 Analysis for Data Traffic

3.4.1 RA Traffic Model

In this section, we analyze throughputs and average delay of RA traffic under the following model. Infinite data stations transmit data packets using asynchronous non-persistent CSMA/CD protocol on an error-free bus. A data station senses the channel even when it has nothing to transmit. When a packet is generated at a data station, the station immediately transmits the packet if the channel is sensed idle. Otherwise, the station retries the transmission of the packet after a randomly distributed retransmission delay.

When a station detects that its transmitting packet was collided, it immediately sends jam signal for a time of T_j and terminates its transmission. The transmission of the collided packet is rescheduled according to the random delay distribution. All stations that detect a jam signal do not transmit any packet for a predetermined time interval called slot time in [IEEE85a], in order to pass a packet fragment generated by a collision. All stations are assumed to behave as if there are carrier on the channel during slot time.

A random access subframe (RF) is modeled as having alternating busy periods and idle periods as shown in Fig.3.2. Busy periods have two states; collision period (BC) and successful transmission period (BT). In BC, all packets transmitted during this period cause a collision and eventually their transmissions are failed. In BT, actually one packet is successfully transmitted.

We characterize the traffic as follows. Data packets are generated in a Poisson stream with parameter λ. Offered traffic load of RA packets including rescheduled ones is assumed to obey a Poisson stream with parameter G_d. And we let L_d and $f_d(x)$ be the average length of data packet to be transmitted and the probability density function (p.d.f.) of the packet length, respectively.

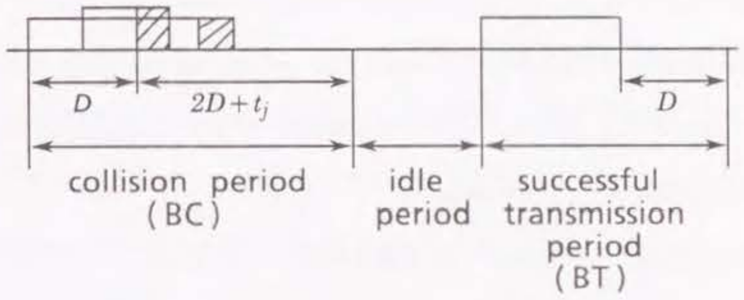


Fig. 3.2 Random Access Subframe (RF) : busy and idle periods

3.4.2 Analysis

At first we will derive the probability that n packets are successfully transmitted during a DF on the condition of DF length being y, $P_{th}(n|y)$. To facilitate the presentation, the following notations are used in this section. It should be noted that, in the remainder of this section, we focus the situation during one VF, and then the time axis is initialized at the beginning of a VF.

- $b_s(t)$: probability density function (p.d.f.) of successful transmission period,
- $b_c(t)$: p.d.f. of collision period,
- $i(t)$: p.d.f. of idle period,
- b_{smin} : minimum time for successful transmission period
- $mc_n(t)$: maximum number of pieces for collision periods which may be occurred by time t on condition that n packets have been successfully transmitted.
- $ns(t)$: maximum number of packets which may be successfully transmitted by t.
- P_s : probability that a packet is successfully transmitted during a busy period.

Since BT is the sum of the packet transmission time and a propagation delay, $b_s(t)$ is given by

$$b_s(t) = \begin{cases} f_d((t-D)\,C) & (\text{if } t \geqq D) \\ 0 & (\text{if } t<D) \end{cases} \tag{3.15}$$

BC is the sum of the time that a station detects a collision and the time to pass a packet fragment generated by a collision. Then, $b_c(t)$ is described as

$$b_c(t) = \delta(t-3D-t_j)\ . \tag{3.16}$$

where, $\delta(t)$ is the Dirac δ-function.

$i(t)$, P_s, $mc_n(t)$ and $ns(t)$ are described as follows,

$$i(t) = G_d\, exp(-G_d\, t) \tag{3.17}$$

$$P_s = exp(-D\cdot G_d) \tag{3.18}$$

$$mc_n(t) = [(t - n\cdot b_{smin})\,/\,(3D+t_j)] \tag{3.19}$$

$$ns(t) = [t\,/\,b_{smin}] \tag{3.20}$$

where, $[x]$ is the Gaussian symbol and means the maximum integer which does not exceed x.

Let $a_{n,m}(t)$ be the probability that at least n packets are successfully transmitted by t on condition that m pieces of collision periods are occurred before n-th successful transmission. $a_{n,m}(t)$ is described using the p.d.f. $d_{n,m}(t)$ that $m+n$ pieces of busy periods are terminated on condition that the busy periods include m pieces of BTs and n pieces of BCs.

$$a_{n,m}(t) = \binom{n+m-1}{n-1} P_s^{\ n}(1-P_s)^m \int_0^t d_{n,m}(\tau)\, d\tau\ . \tag{3.21}$$

If we denote the Laplace transform of $d_{n,m}(t)$, $b_s(t)$, $b_c(t)$ and $i(t)$, as $d_{n,m}{}^*(s)$, $b_s{}^*(s)$, $b_c{}^*(s)$ and $i^*(s)$, respectively, we have

$$d_{n,m}{}^*(s) = (b_s{}^*(s))^n (b_c{}^*(s))^m (i^*(s))^{n+m}\ . \tag{3.22}$$

The derivations of $a_{n,m}(t)$ and b_{smin} in the case of constant packet length, are described in appendix 3.A.

Let $A_n(t)$ and $F_n(t)$ be the probabilities that at least n packets have been successfully transmitted by t, and that just n packets have been successfully transmitted by t, respectively, then both probabilities are given by

$$A_n(t) = \sum_{m=0}^{mc_n(t)} a_{n,m}(t) \tag{3.23}$$

$$F_n(t) = \begin{cases} A_n(t) - A_{n+1}(t) & (\text{ if } 0 \leqq n \leqq ns(t)) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.24}$$

The RA station should have terminated the packet transmission by the time $y-D$, since the TMC transmits a VFS signal every frame time independently on the status of channel. Consequently, the probability $P_{th}(n|y)$ that n packets are successfully transmitted during a RF, on condition that the length of RF is equal to y, is described by

$$P_{th}(n|y) = F_n(y-D) \tag{3.25}$$

The RF length is distributed discretely, since the VF length is distributed discretely and the sum of VF length and RF length in a frame is constant. Then, we denote,

Y : the set of available RF length y,

$P_{ld}(y)$: probability that the RF length is equal to y,

$P_{ds}(n)$: probability that n RA packets are successfully transmitted in a frame.

$P_{ds}(n)$ is then

$$P_{ds}(n) = \sum_{y \in Y} P_{ld}(y) \cdot P_{th}(n|y) \tag{3.26}$$

The throughput of RA traffic S_d is defined as the average number of successfully transmitted packets per unit time, and is given by

$$S_d = \sum_{n=0}^{ns(y-D)} n \cdot P_{ds}(n) / T_f \tag{3.27}$$

If we assume the back off time of collided packets are exponentially distributed with mean t_b, and offered traffic load G_v is also exponentially distributed. Then we approximately obtain the average delay of RA packets $E[D_d]$ as

$$E[D_d] = t_b (G_d / S_d - 1) \ . \tag{3.28}$$

In what follows we describe the $P_{ds}(n)$ for both proposed protocols.

[Average delay of RA packets for M-TDM protocol]

Let t_{mc} be the sum of the time intervals of VFS and VFE signals, and $y(i)$ be the RF length when n VC stations are in ready mode. Then, $y(i)$ is given using eq.(3.9) by

$$y(i) = T_f - x(i) + t_{mc} \ . \tag{3.29}$$

Subsequently, $P_{ds}(n)$ is derived from eqs.(3.26) and (3.29), as

$$P_{ds}(n) = \sum_{i=0}^{N_v} p_i \cdot P_{th}(n|y(i)) \ . \tag{3.30}$$

[Average delay of RA packets for M-ATDM protocol]

Let $y(i, j)$ be the RF length when j of i ready stations transmit actual packet during a VF. $y(i, j)$ is given using eq.(3.12) by

$$y(i,j) = T_f - x(i,j) - t_{mc} \ . \tag{3.31}$$

Subsequently, $P_{ds}(n)$ is derived from eqs.(3.26) and (3.30), as below.

$$P_{ds}(n) = \sum_{i=0}^{N_v} \sum_{j=0}^{i} p_i \cdot P(i,j) \cdot P_{th}(n|y(i,j)) \ . \tag{3.32}$$

3.5 Numerical Examples and Discussions

In the following examples, we consider the case that the transmission capacity of the channel C is 10 Mbit/s, propagation delay D is 20 μsec, the length of a frame T_f is 10 msec, the length of jam signal is 32 bits (i.e., $t_j=32/C$ sec), the length of RA packet L_d is fixed to 3000 bits, the length of VC packets is 784 bits, the length of short packets H is 144 bits, the average holding time of a VC call 1/μ is 100 sec, the probability of packet generation during a call α is 0.366, and time interval of the sum of VFS and VFC, t_{mc} is 60 μ sec.

At first, we will examine the validity of access protocol of a new VC connection. Table 3.1 shows the throughput of VC traffic S_v and the probability of the call rejection for various VC access rate G_v, where, the maximum number of VC is assumed to be 70. The probability that the access is failed due to collision is vary small. It means that if the VC is not full, Almost all of the calls will succeed in setting up a circuit with one or at most few trials.

In the following examples, we also show the results of fixed boundary model (F-TDM) for comparison. Fig. 3.3 shows the average length of virtual circuit subframe (VF) as a function of offered load of RA calls G_v. For both movable protocols, the VF lengths change according to the offered load. It means that our proposed protocols make an efficient channel use.

Next, we consider the performances of RA traffics. In Figs. 3.4 and 3.5, lines and symbols show the analytical results and simulations, respectively. In the analytical model we assume the infinite RA stations, whereas, in the simulations, we assume more practical models that the number of the RA stations is 800, and that the buffer size of each station is either one or infinite. It should be noted that the simulation was employed for F-TDM model, but not for movable boundary model, since the behavior of movable boundary models are very complicated and the simulations for them take a lot of calculation time. In order to show the accuracy of the analysis, we compare the analytical

Table 3.1 Throughput and reject probability of VC call.

G_v (packet/ms)	S_v (packet/ms)	P_{fault}	P_{full}	P_{coll}
0.20	0.1996	0.0020	0.0000	0.0020
0.40	0.3984	0.0040	0.0000	0.0040
0.60	0.5833	0.0279	0.0221	0.0060
0.80	0.6614	0.1732	0.1675	0.0080
1.00	0.6888	0.3202	0.3134	0.0100
1.50	0.6914	0.5391	0.5321	0.0149
2.00	0.6946	0.6527	0.6457	0.0198
($\times 10^{-4}$)	($\times 10^{-4}$)			

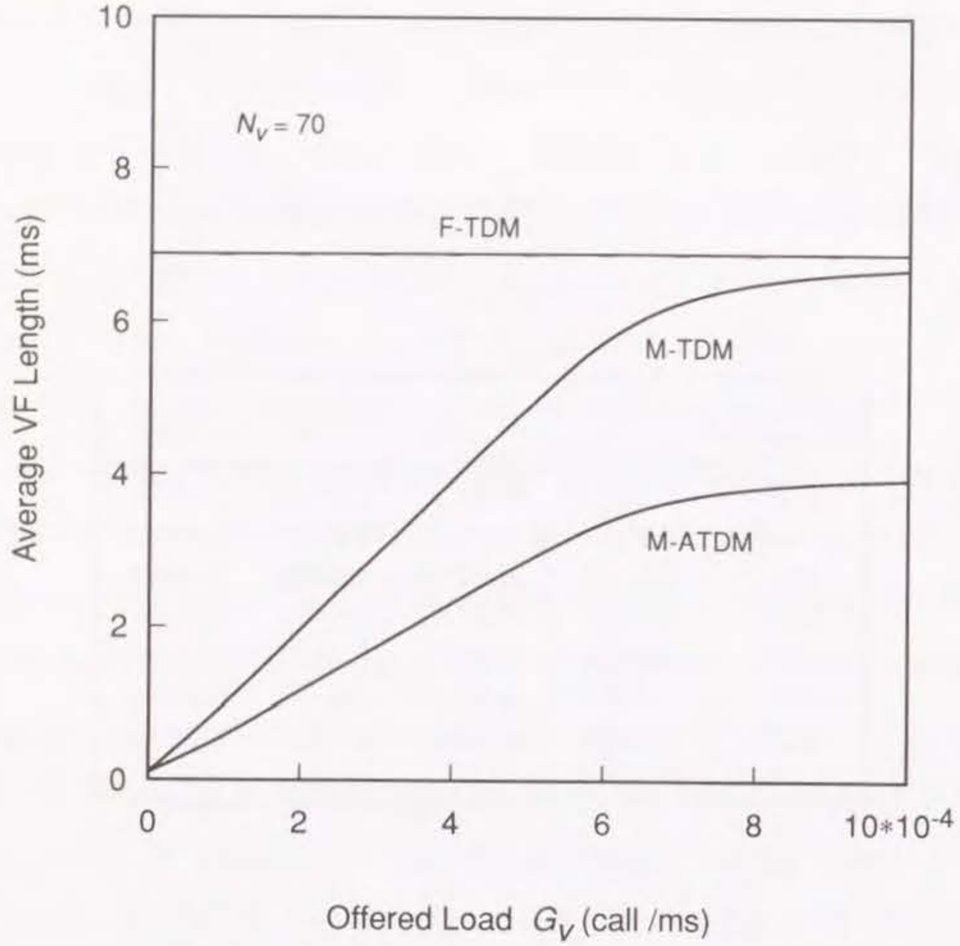


Fig. 3.3 Average length of VC subframe vs. offered load of VC call.

results with simulations in Fig. 3.4, where throughput of RA traffics for F-TDM model are plotted as a function of offered traffic load of RA packets. Analytical results match the simulations very well, especially, for the larger back-off time t_b, analytical results plot almost the same points.

In Fig. 3.5, we consider the validity of the choice of back-off protocol, (i.e., no-persistent CSMA/CD). The results are compared with those for 1-persistent CSMA/CD protocol. For both types of protocols, each RA station is assumed to have single buffer. Fig. 3.5 shows the average delay of RA packet as a function of offered load. Maximum throughput of non-persistent models are superior to that of 1-persistent model. Especially, in the case for $t_b=2$, non-persistent model has the smaller delay for large G_d. From this figure, we can find that non-persistent back off is suitable for Hybrid VC/CSMA-CD protocol.

Fig. 3.6 shows throughput of RA packets as a function of offered load of RA packets, where, $G_v=3.0\times10^{-4}$ and $N_v=50$. For all ranges of G_d, M-ATDM plots the highest throughput, and F-TDM plots the worst throughput. M-ATDM protocol can transmit the RA packet two times as much as F-TDM protocol. Fig. 3.7 shows the average delay of RA packets for the same situation as Fig. 3.6. If the system is not stable, the average delay is less than 20 msec for $t_b=2$, and less than 100 msec for tb=10. It means that if the RA traffic is restricted to the stable level, our protocols satisfy the requirement for the delay of RA packets.

Figs. 3.8 and 3.9 show the maximum throughput of RA packets and average VF length respectively, as a function of maximum allowable number of VCs for constant load of G_v. Since our protocols give variety to the VF length according to the VC load, they prevent the system from wasting the channel like as M-TDM model.

From these results, we can conclude that the protocols proposed in this chapter perform well for both VC and RA traffic. M-ATDM achieves the highest performance among three protocols compared in this section. On the contrary, M-TDM produce the similar performance with the simple control mechanism.

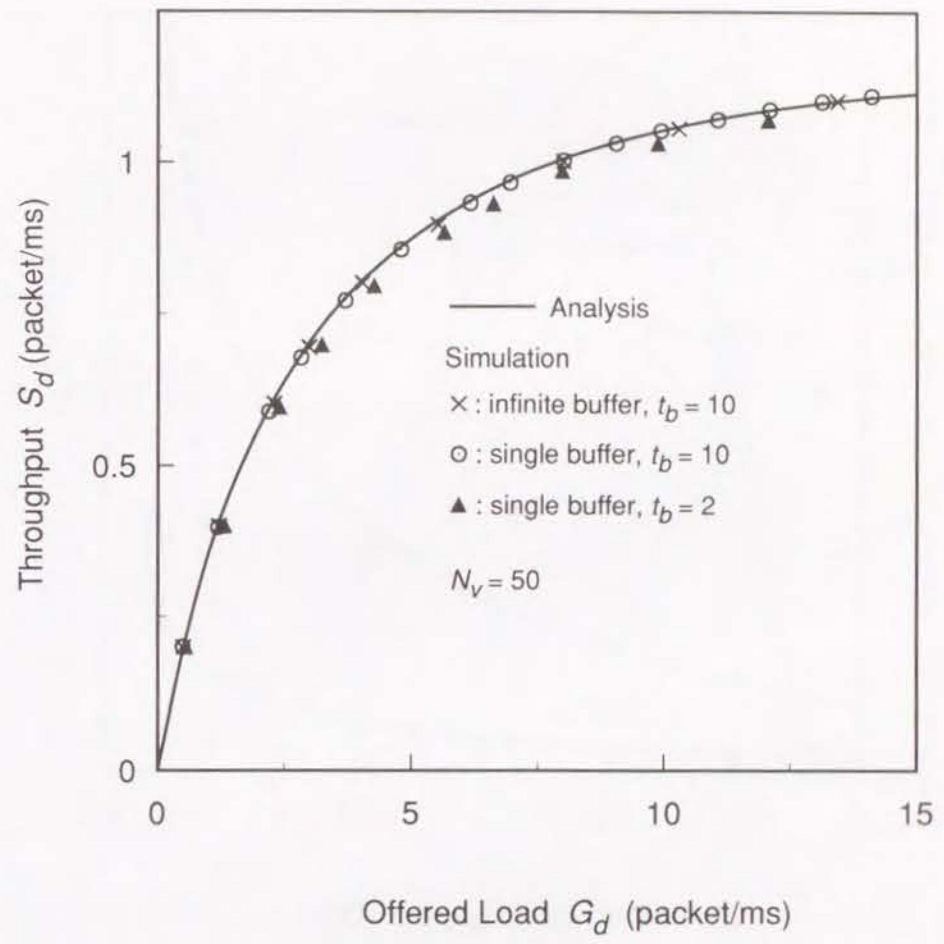


Fig. 3.4 Offered load vs. throughput of RA packet for fixed boundary model.

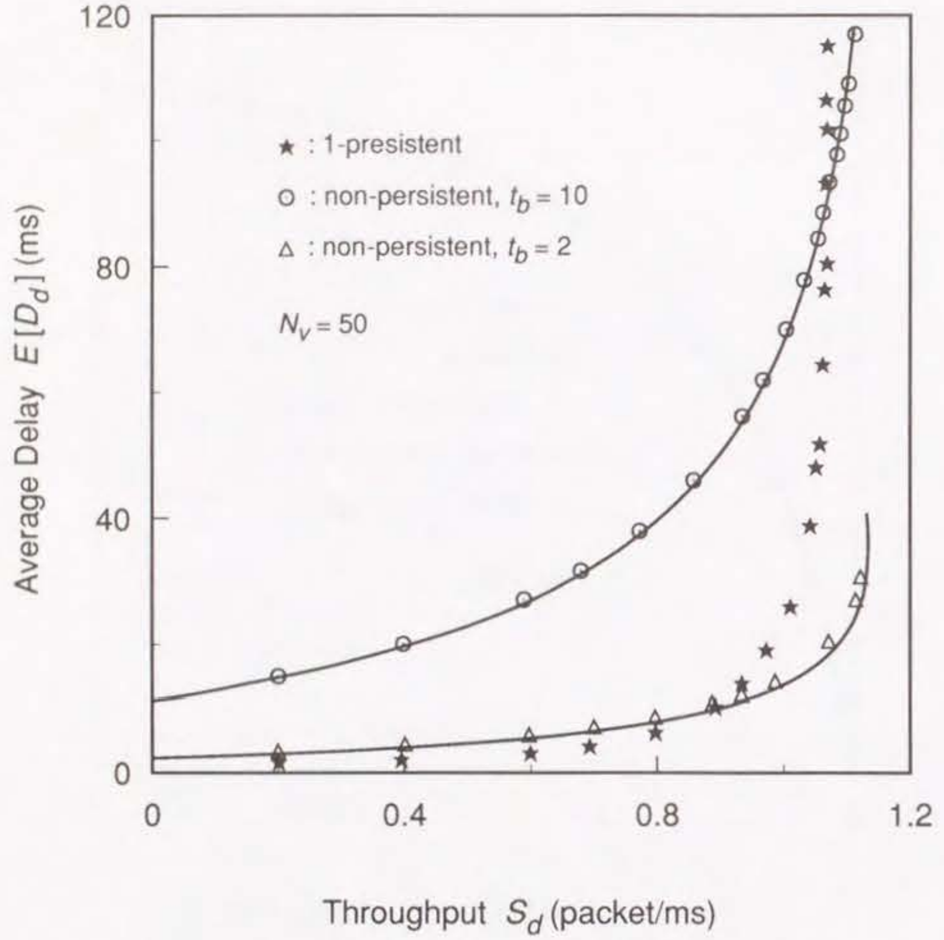


Fig. 3.5 Average delay vs. throughput of RA packet for fixed boundary model.

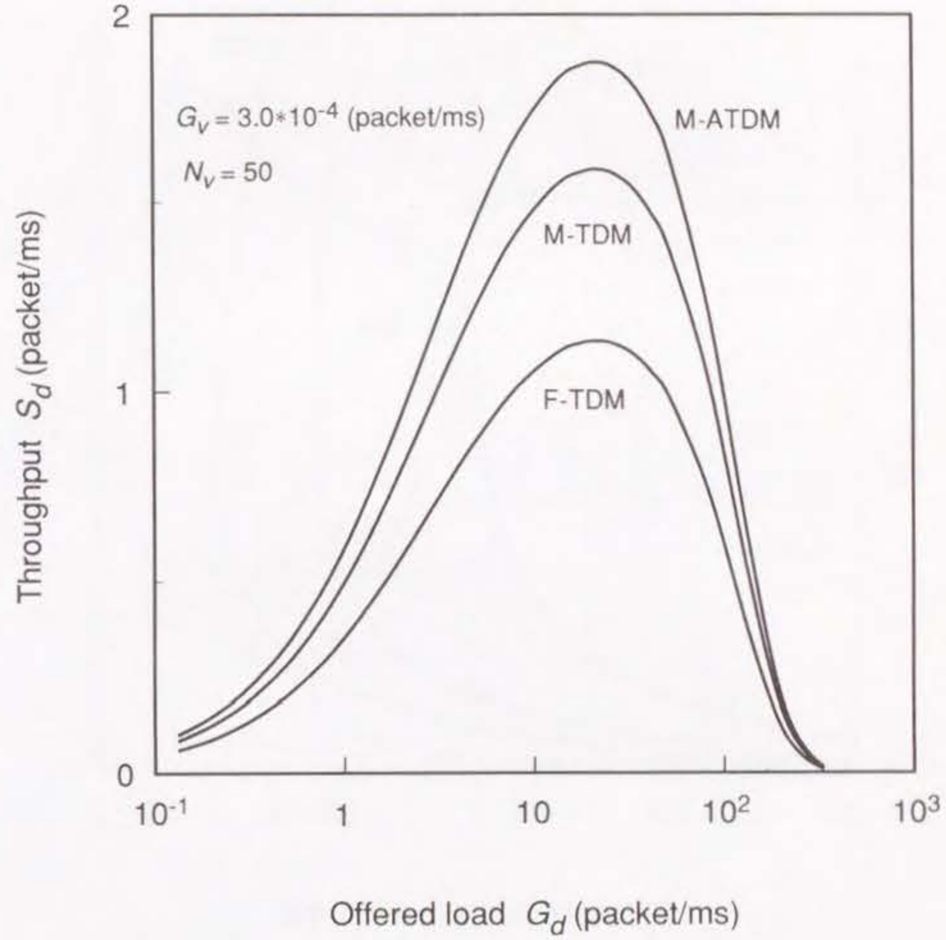


Fig. 3.6 Throughput vs. offered load of RA packet.

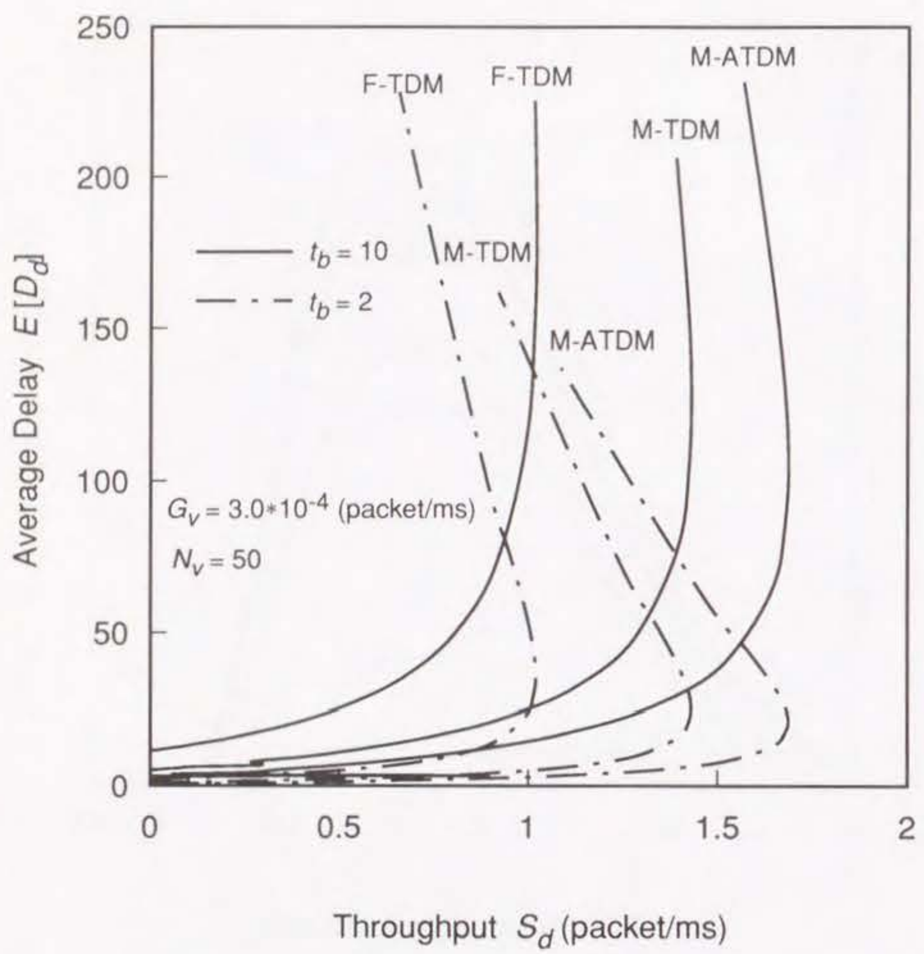


Fig. 3.7 Average delay vs. throughput of RA packet.

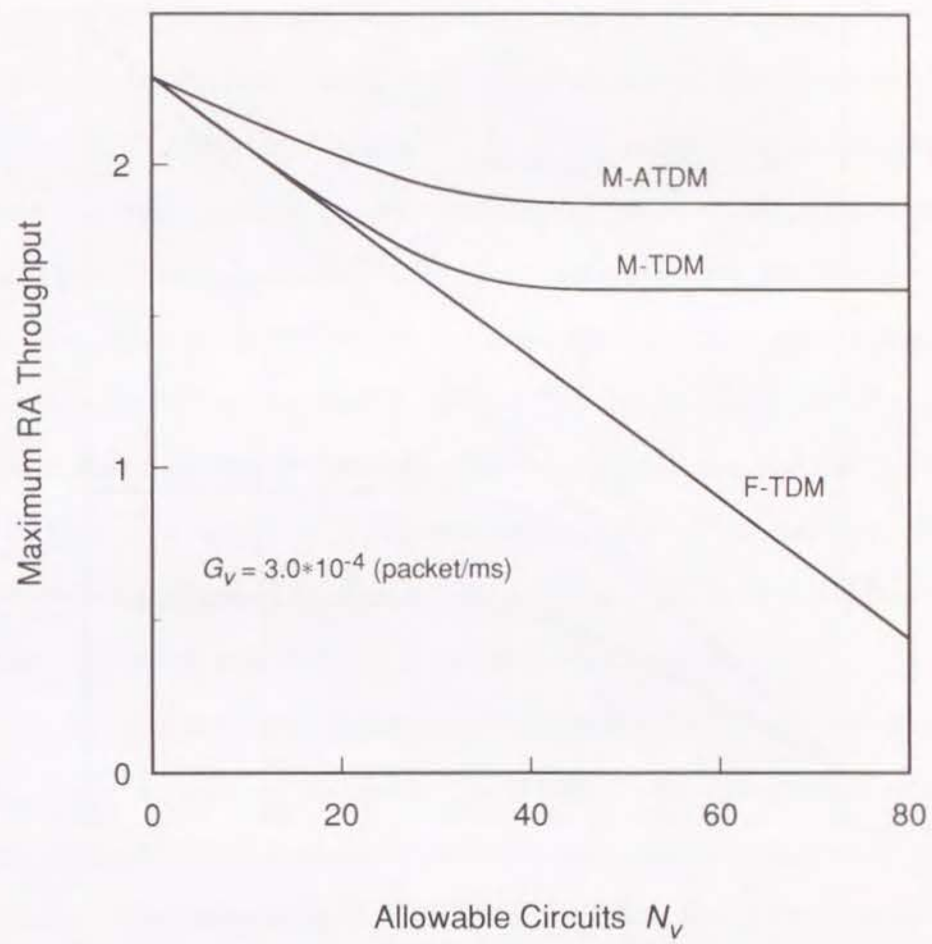


Fig. 3.8 Maximum throughput of RA packet vs. number of allowable virtual circuits.

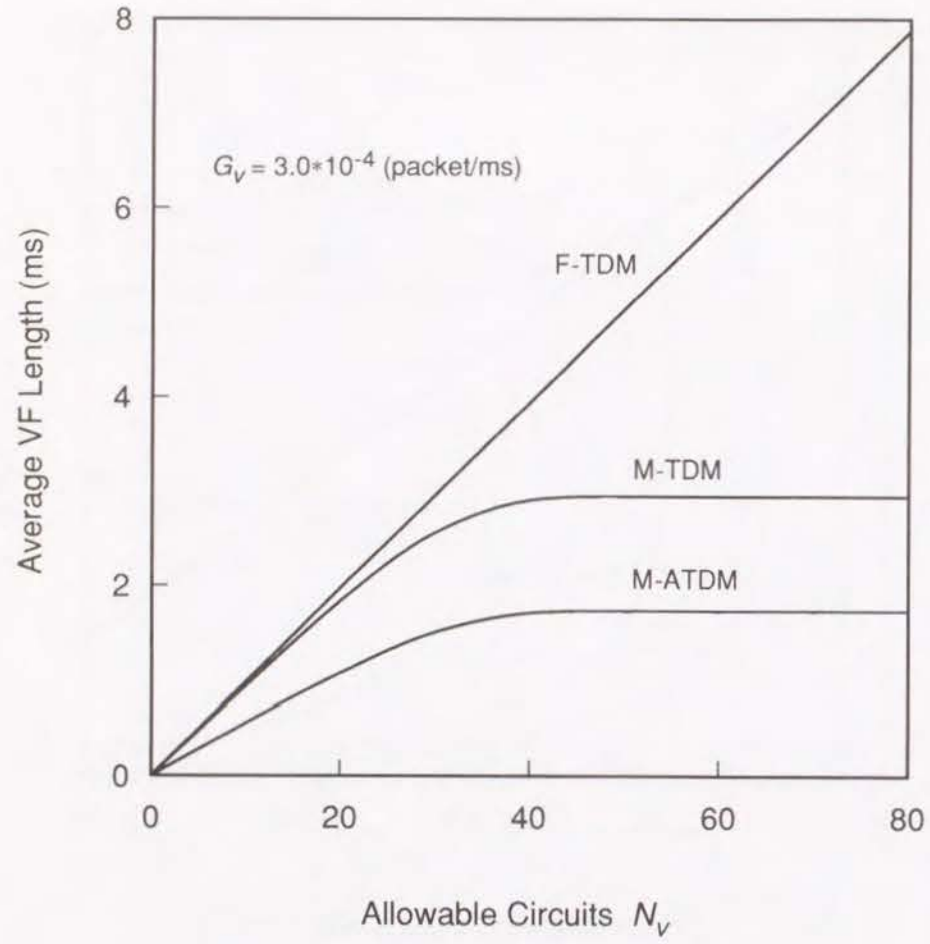


Fig. 3.9 Average length of VC subframe vs. number of allowable virtual circuits.

3.6 Conclusion

A hybrid VC-CSMA/CD protocol in integrated local area bus networks has been proposed and analytically evaluated. A fixed length frame consists of a virtual circuit (VC) subframe and a random access (RA) subframe, which are separated with movable boundary. In a VC subframe, connection oriented stations transmit packets in a virtual circuit fashion, which ensures a collision free and real-time packet delivery. In an RA subframe, random access stations access the channel according to the asynchronous non-persistent CSMA/CD protocol. Two effective and practical VC techniques were proposed and analyzed using queueing models. The analytical method for the throughput and delay of RA traffic and the loss probability of VC call was derived. The analytical and simulation results showed that the protocol presented in this chapter performed well for both types of traffic.

In the future network systems, integrated LANs will be connected each other, and also will be connected to global area integrated networks. To establish the efficient gateway protocol which connect integrated LANs and to evaluate the performances of interconnected integrated systems are the future work of interest.

Appendix 3.A Derivation of $a_{n,m}(t)$ and B_{smin}

We consider the case of constant packet length. Since the time interval of BT period is the sum of a packet transmission time and a propagation time, eq. (3.15) is rewritten using Dirac's δ-function as below.

$$b_s(t) = \delta(t-(L_d/C+D)) \tag{3A.1}$$

Laplace transform $b_s^*(s)$, $b_c^*(s)$ and $i^*(s)$are obtained from eqs. (A3.1), (3.16) and (3.17), respectively.

$$b_s^*(s) = exp(-(L_d/C+D)s) \tag{3A.2}$$

$$b_c^*(s) = exp(-(3D+t_j)s) \tag{3A.3}$$

$$i^*(s)=\left(\frac{G_d}{G_d+s}\right). \tag{3A.4}$$

Subsequently, using the inverse Laplace transform (L^{-1}), $d_{n,m}(t)$ is given by

$$d_{n,m}(t) = L^{-1}d_{n,m}^*(s) = L^{-1}(b_s^*(s))^n(b_c^*(s))^m(i^*(s))^{n+m}$$

$$= L^{-1}(exp(-(L_d/C+D)s))^n(exp(-(3D+t_j)s))^m\left(\frac{G_d}{G_d+s}\right)^{n+m}$$

$$= \frac{G_d^{\,n+m}}{(n+m-1)!}(t-b_{m,n})^{n+m-1}exp(-G_d(t-b_{m,n}))\,\theta(t-b_{m,n}) \tag{3A.5}$$

where

$$b_{m,n} = n\cdot(L_d/C+D) + m(3D+t_j) \tag{3A.6}$$

$$\theta(t) = \begin{cases} 1 & (\text{if } t\geqq 0) \\ 0 & (\text{if } t<0) \end{cases}. \tag{3A.7}$$

Accordingly, we have

$$\int_0^t d_{n,m}(\tau)\,d\tau = \int_0^{t-b_{m,n}} \frac{G_d^{\,n+m}}{(n+m-1)!}\tau^{n+m-1}exp(-G_d\tau)\,d\tau$$

$$= 1-exp(-G_d(t-b_{m,n}))\sum_{r=0}^{n+m}\frac{(G_d(t-b_{m,n}))^r}{r!}. \tag{3A.8}$$

By substituting eq.(3A.8) in eq. (3.22), we obtain $a_{m,n}(t)$.

Since we assume the constant packet length, the minimum BT period B_{smin} is equal to the sum of the packet transmission time and propagation delay, and is given by

$$B_{smin} = L_d/C+D. \tag{3A.9}$$

CHAPTER 4

Performance Evaluation of Packetized Voice Transmission on a Token Passing Ring Network

4.1 Introduction

In the past few years, local area networks for interconnection of computers and shared resources within a small area such as a building or a campus have rapidly gained in popularity. Many architectures have been proposed for local networks, and several have been implemented. The token passing ring network is one of the most popular and prominent local networks. Various studies on its performance have proved its merit for a variety of data transfer applications. As a rule of thumb, a token passing ring network is less sensitive to a change in a network load, offers reasonably short delay under light load, and carries more traffic under heavy load. Furthermore, it gives less variance in transmission delays than contention type networks such as Ethernet, where large delay variance is unavoidable due to the undeterministic nature of contention and collision/backoff [BUX81].

Recently, interest has focused on the use of token passing ring networks for real-time voice applications to provide both computer and telephone communication support to network users. The traffic characteristics and performance requirements for voice and data are quite dissimilar [COHE81, GOPA81]. Data traffic can be categorized into two basic types; interactive and bulk data. Interactive data traffic is bursty in nature; the actual proportion of bandwidth utilized is typically very small. It consists of short messages requiring network delay. Bulk data, on the other hand, consists of long messages, and requires high throughput, but real-time delivery consideration is not of its primary importance. Both types of data are random in generation. Strict error control and recovery procedures are required for both types of data.

Voice calls last for a few minutes, with for more than 60% of the time the channel remains idle. This is due to the fact that the only one speaker is active at any time; furthermore, there are pauses between sentences, phrases, and even between syllables. Voice traffic can be modeled as having alternating talkspurts and silences, with generation of voice packets at a constant rate during talkspurts and no packet at a constant rate during talkspurts and no packet generation during silence periods. For the integrity of the conversation, voice packets must be delivered within some bound (typically 50-100 msec); but in many data applications a delay of as much as 100 or 200 msec does not present a problem. Conversation is inherently robust, and as a result, speech can be reconstructed at the destination with acceptable quality, provided that the information loss is less than some specified fraction (typically 1%).

Despite its importance, little has been known about the performance of a token passing ring network when it is subjected to a voice load in addition to a data load. Due to the real time constraints of interactive voice applications, it is not apparent to us if the token passing ring network offers a viable alternative to other methods of interconnecting users. The objectives of this chapter are to evaluate the performance of a token passing ring network through simulations and to explore the feasibility of this class of network as a medium for combined voice and data transmission[SUDA85]. Performance measures obtained include the distribution of the queueing delay, loss probability of voice packets, and the number of users allowed on a network to satisfy real-time constraints of speech.

In section 4.2, we describe the network protocol in detail. Traffic model assumed in our simulations is discussed in section 4.3, followed by discussion on the performance measures for packetized voice networks in section 4.4. Numerical results are presented in section 4.5.

4.2 Network Model

Two types of users are assumed on a network: voice users and data users. At a voice user, a continuous voice analog signal is digitized by coder (Fig. 4.1). For instance, a typical PCM encoder operating at 8 KHz produces one 8-bit word every 125 micro sec. Generated samples are accumulated in the packetizer. When samples in the packetizer reaches the (predetermined) packet length, a header is attached and a voice packet is generated. A generated voice packet is then examined by a speech activity detector if it contains enough speech activity. Silent packets are discarded, and non-silent packets are only stored in the buffer (in the order of generation) and wait for their transmission. Packet generation cycle is independent of the packet transmission process, and the queue size at the buffer continues to grow while a packet is waiting for transmission. Since we assume infinite buffer capacity, there is no packet loss at a voice source user.

Transmission control is based on a token, which is passed from a user to a user around the ring in the predetermined sequence. No priority is assumed on voice users over data users. Both types of users follow the same transmission scheme. When a free token is passed to a user ready to send a packet, the user changes the token to a busy status and then appends his packet. We assume that the packet at the head of the buffer is only served by a token. The rest in the buffer has to wait for the next (and successive) visit of a free token. In order to minimize the loss of voice packets due to excessive transmission delay, variance in transmission delays should be kept small. It is intuitively clear that this service discipline gives less variance in delays than other disciplines such as exhaustive service, where the token serves each user until its buffer is emptied. The transmitted packet circulates around the ring, to the intended destination user, and then returns to the transmitting user to acknowledge

complete successful transmission around the ring. The token is then released to the next user downstream with packets.

No specific play-out strategy is assumed at destination users. A user plays out voice packets one by one upon their arrivals. Infinite buffer space for storing incoming voice packets are assumed, and hence, there is no packet loss at the destination due to buffer overflow.

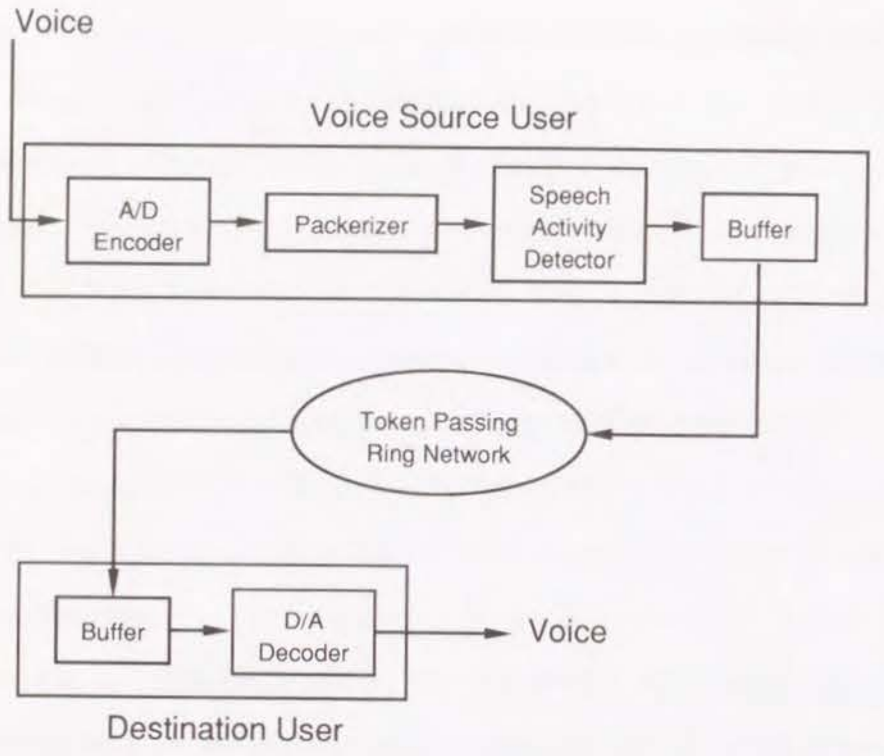


Fig. 4.1 Voice users.

4.3 Voice and Data Traffic Models

In our simulations, both voice and data traffic are modeled. The data packet arrivals are modeled by a Poisson process. The traffic generated by a Poisson process is not as bursty as real data traffic, but the difference is expected to be relatively unimportant in determining the effect of the data traffic upon voice traffic.

It has been well known that arrival process of new voice calls and the distribution of their durations can be characterized by the Poisson process and an exponential distribution, respectively[COHE81]. However, termination of calls and generation of new calls are not considered in our simulations. All voice users are assumed to be active with calls throughout the simulation period. The reason is that the statistical fluctuations in the presence of talkers are much slower than the statistical fluctuations in the generation and transmission of voice and data packets. The holding time per voice calls is on the order of hundreds of seconds while the holding time for voice and data packets are on the order of tens of milliseconds (or less).

Since we assume consistent active voice users, voice traffic at a user is modeled as having alternating talkspurts and silences, with generation of voice packets at a constant rate during talkspurts and no packet generation during silence periods (Fig. 4.2). If we denote

V : voice coding rate (bit/sec),

P_v : voice packet length excluding header (bit) ,and

H : voice packet header length (bits),

voice packet generation period W_p becomes P_v/V sec. W_p is the delay introduced by accumulating enough samples to construct a voice packet.

The statistics of the duration of talkspurts and silences in conversational speech depend on mechanism to detect speech activity. For instance, the average lengths of talkspurts and silences are assumed to be 1.23 sec and 1.77 sec in [BRAD65], 0.185 sec and 1.31 sec in [MAXE82], and 0.17 sec and 0.41 sec in [GRUB82b], respectively. The effect of different talkspurt/silence statistics on the network performance is also examined in this chapter.

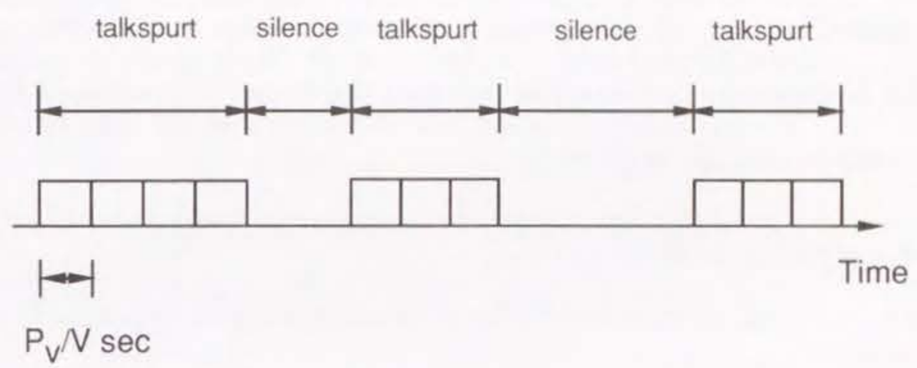


Fig. 4.2 Voice packet arrival process.

4.4 Performance Measures

Since excessive delays can have serious disruptive effects on human conversation, voice packets must be received at the destination user, within given amount α of time after their generation at the source user. Those packets that do not arrive within this time bound are considered lost; a small number of lost packets have been shown to have little, if any, effect on human speech intelligibility.

Transmission delay D_v of a voice packet is defined as the time interval between the beginning of its digitization and the time it is removed from the ring by the source user. D_v is given by

$$D_v = W_p + W_q{}^v + T_v + R + B \tag{4.1}$$

where

$W_q{}^v$: queueing delay of a voice packet,

T_v : transmission time of a voice packet,

R : propagation delay of a ring, and

B : total bit latency of a ring.

W_p is the voice packet generation period, and this period depends on the packet size chosen. Further delay is incurred in gaining access to the network (W) and in the actual transmission of the packet (T_v). B is the total bit latency of the network. This is the delay introduced at users to monitor and change the token bit pattern. 1-bit latency is assumed at each user in our simulations. For each of the above delay elements, we have

$$W_p = P_v / V \tag{4.2}$$

$$T_v = (P_v + H) / C \tag{4.3}$$

$$R = UL \tag{4.4}$$

$$B = N / C \tag{4.5}$$

where

C : channel speed (bit/sec)

L : physical length of a ring

U : media propagation delay per unit cable length

N : the total number users on a ring

N is the sun of the numbers of voice users, N_v and data users N_d.

Since packets that do not arrive within the given time bound α are considered lost, the loss probability of voice packets becomes

$$P_{loss} = Prob[D_v > \alpha] = 1 - D_v(\alpha) \tag{4.6}$$

where $D_v(x)$ is the distribution function of voice packet delay D_v, namely, $D_v(x) = Prob[D_v \leqq x]$.

The transmission delay D_d of a data is defined in a similar way and becomes

$$D_d = W_q + T_d + R + B \tag{4.7}$$

where, W_q is the queueing delay of a data packet at the source user, and T_d is the transmission time of a data packet given by $T_d = P_d / C$. P_d is the data packet length (including its header) in bit.

4.5 Simulation Results

4.5.1 Parameter Values

In simulations, we assume that the average lengths of talkspurts (T_{talk}) and silences (T_{sile}) within a call 0.17 sec and 0.41sec, respectively [GRUB82] (except in Fig. 4.7). PCM voice coding scheme is assumed at all voice users, i.e., $V=64$ Kbits/sec. Length of the ring (L) is 1 km, and users are distributed uniformly on the ring. Values for the token length (3 octets), packet header length H (21 octets) and channel speed C (1 Mbits/sec)are based on the IEEE standard 802.5 for a token ring local area network [IEEE85b]. Propagation delay is assumed to be 5 μsec per km of cable. It should be noted that our assumption of only 5 μsec propagation delay per km cable represents an optimistic view, since other delay component such as signal rise times or repeater delays may significantly increase the total end-to-end propagation delay[BUX81].

4.5.2 Numerical Examples: Voice Users

We first show simulation results for the cases where there are only voice users on a network.

Fig. 4.3 shows the average delay $E[D_v]$ of voice packets along with its components W_p, $E[W_q{}^v]$ (the average queueing delay) and T_v as a function of P_v. The value of $R+B$ is small (35 μsec) compared with the values of other delay components and is not shown in the figure. The number of users N ($=N_v$) on a network is 30 in this figure. $E[D_v]$ and $E[W_q{}^v]$ are obtained through simulations, while W_p and T_v are calculated from eqs. (4.2) and (4.3), respectively. The optimal value P_{opt} for a voice packet length to minimize D_v is 640 bits. This is due to the following tradeoff relation; as P_v increases, so do W_p and T_v; on the contrary, W_q decreases due to less overhead in packet headers.

Figs. 4.4 and 4.5 show the density function of D_v and P_{loss} respectively. In these figures, N_v is assumed to be 30. In Fig. 4.5 voice packet whose delay D_v

exceed 50 msec are considered lost. (We assume 50 msec time bound throughout the chapter.) To satisfy the condition $P_{loss} \leqq 1\%$, voice packet length must be (approximately) within the range of $880 \leqq P_v \leqq 1840$.

Fig. 4.6 shows the average delay $E[D_v]$ and P_{loss} as a function of N_v for various values of P_v. It is shown that the maximum number of voice users N_v allowed on a network to satisfy the condition $P_{loss} \leqq 1\%$ is 28 users, if $P_v=640$ bits, and 30 users, if $P_v=1280$ bits. When P_v is 640 bits, although slightly less number of voice users are allowed on a network, the average delay $E[D_v]$ is much less over the range of our interest than the case with 1280 bit voice packets.

Fig. 4.7 shows the maximum number of voice users $N_v{}^{max}$ allowed on a network to satisfy the condition $P_{loss} \leqq 1\%$ as a function of P_v for different values of (T_{talk}, T_{sile}). Distributions of talkspurts and silences are still assumed to be exponential. For each pair of (T_{talk}, T_{sile}), the ratio, $T_{talk} : T_{sile}$ is kept constant and is equal to 0.17 : 0.41. For the given packet length, $N_v{}^{max}$ decreases as the average talkspurt T_{talk} becomes larger. This is because that longer talkspurt gives larger correlation in voice packet generation.

4.5.3 Numerical Examples : Voice and Data Users

In Figs. 4.8 and 4.9, both voice and data users are assumed on a network. N_v and N_d are 20 and 5, respectively. Data users generate data packets of constant length (1024 bits including a header) according to a Poisson distribution. The average delays of both voice and data packets are shown in Fig. 4.8. Fig. 4.9 shows the loss probability of voice packets. For the given condition $P_{loss} \leqq 1\%$, a network with 1280 bit voice packets supports wider range of data traffic at the expense of larger voice packet delay in light data load.

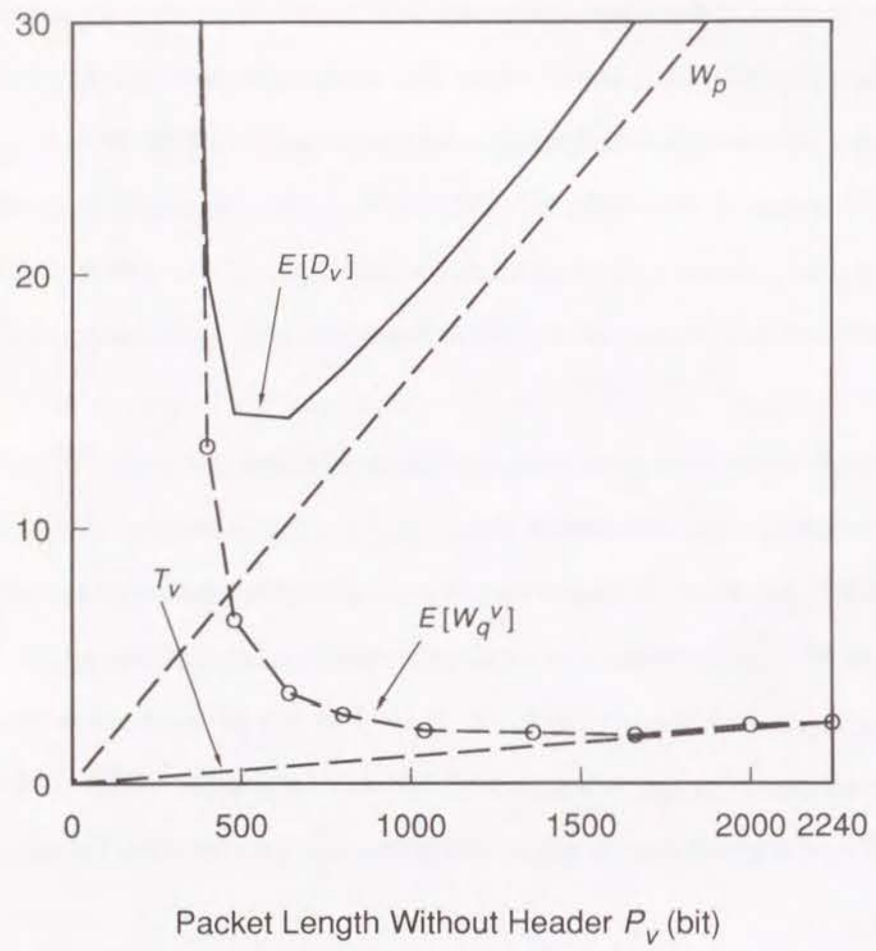


Fig. 4.3 Average transmission delay vs. packet length ($N_v = 30$).

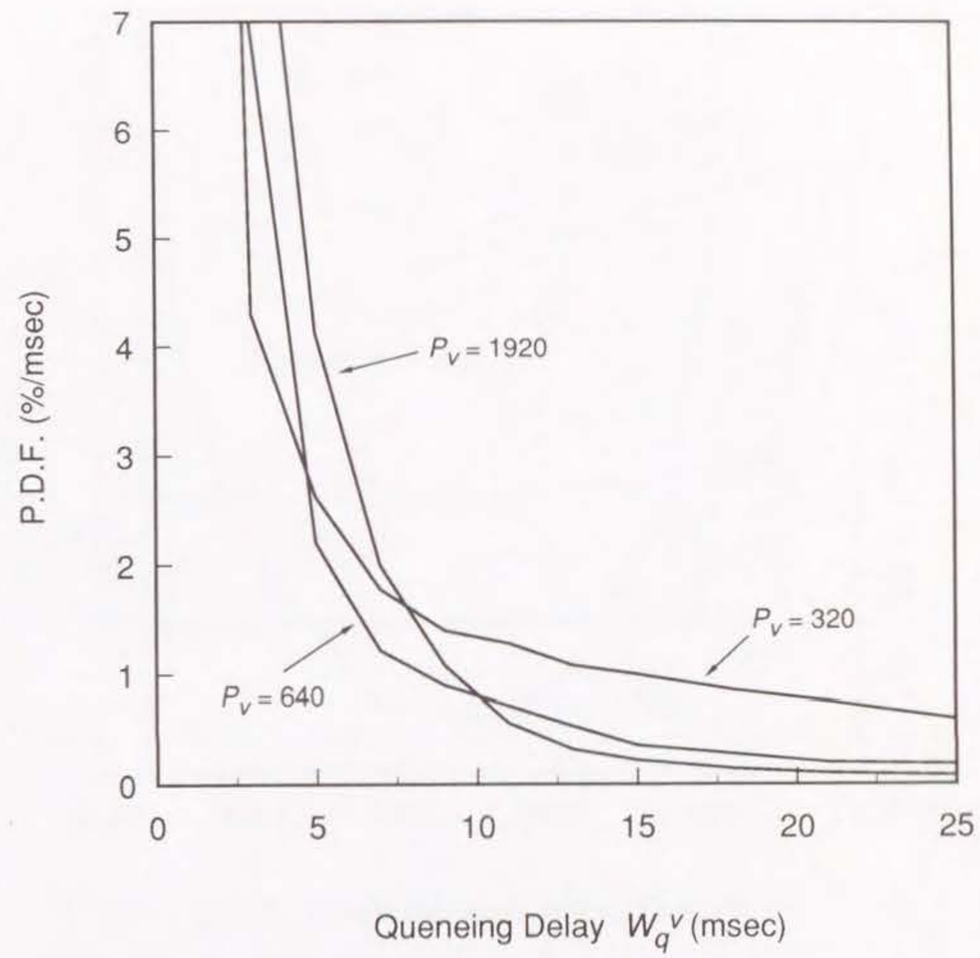


Fig. 4.4 Probability density function of W_q^v ($N_v = 30$).

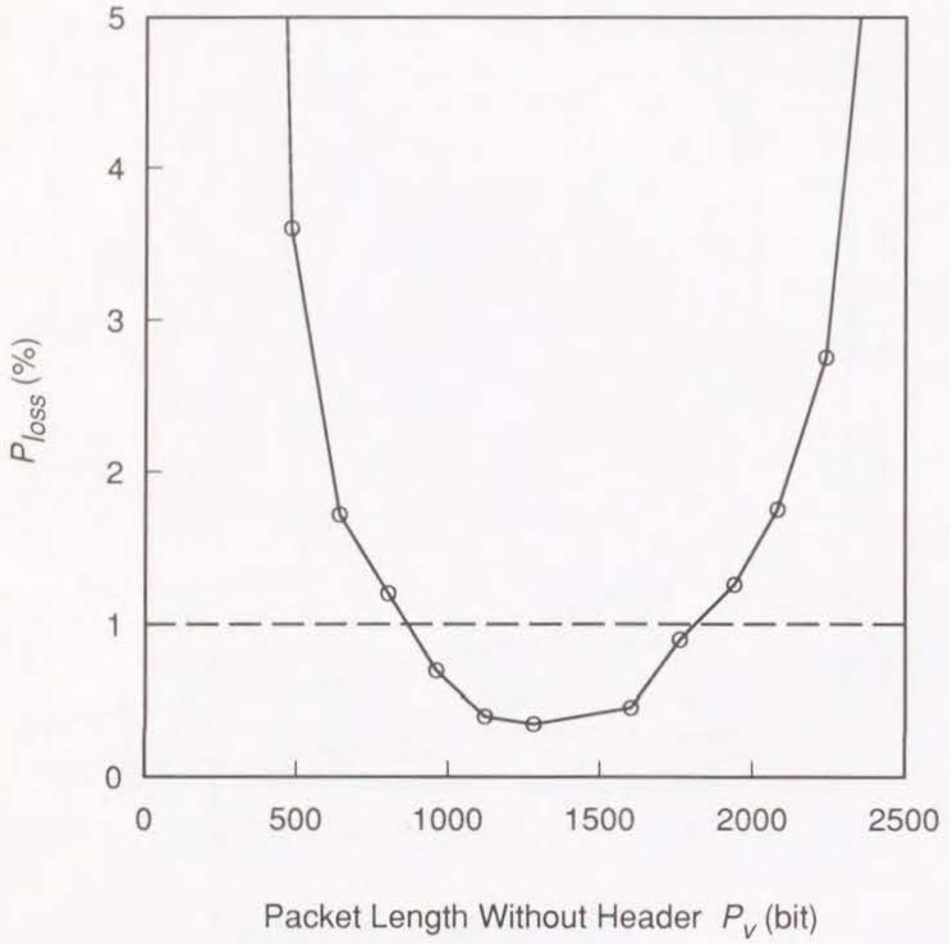


Fig. 4.5 Loss probability vs. packet length.

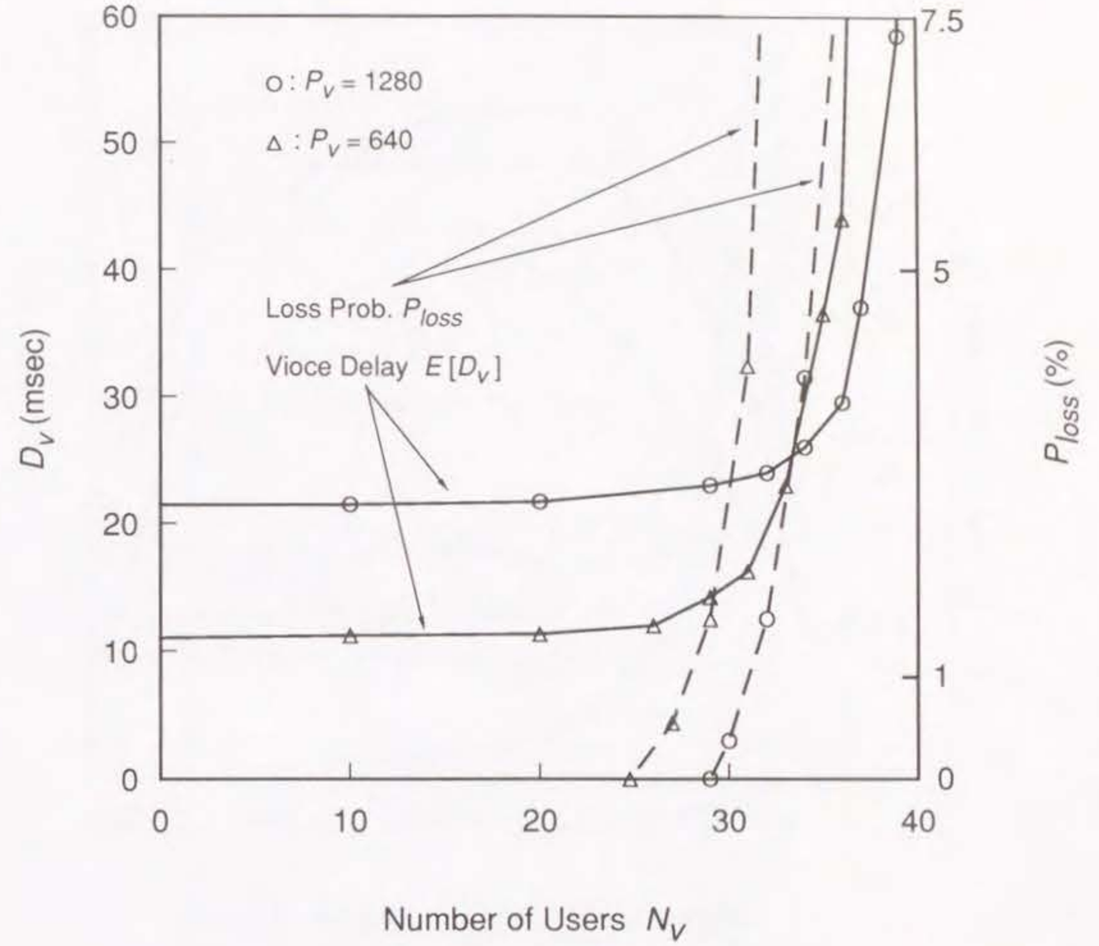


Fig. 4.6 Average transmission delay vs. number of voice users.

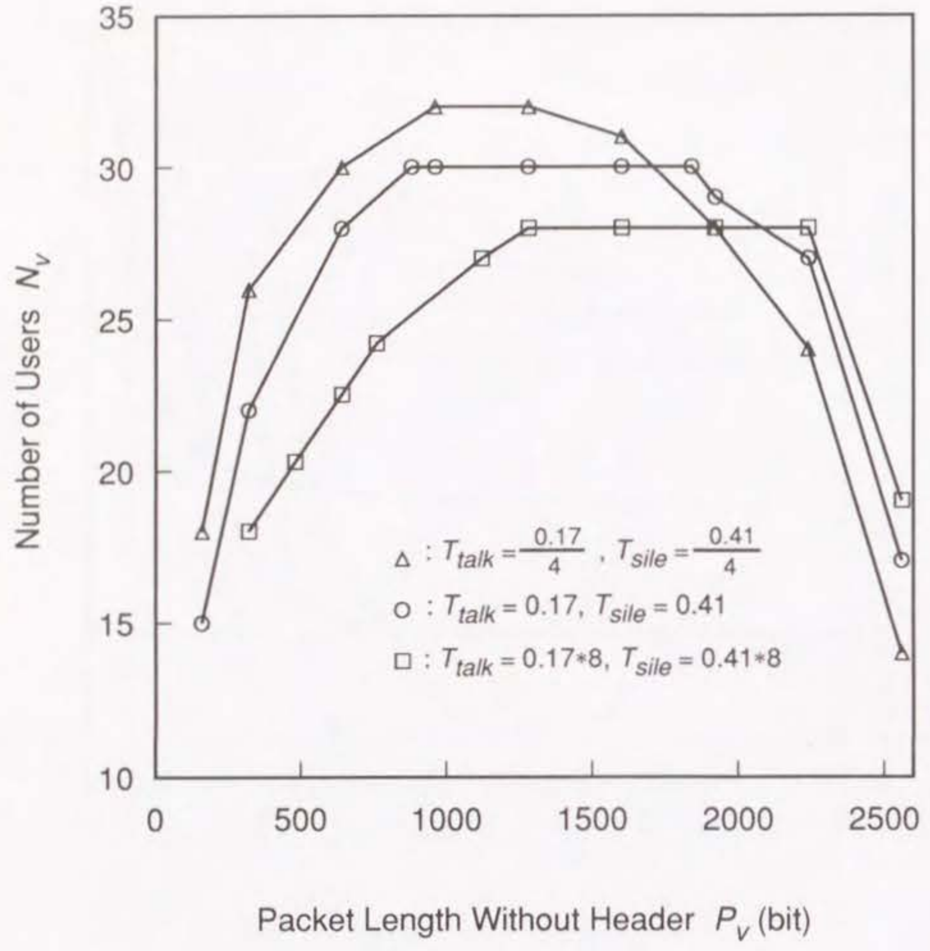


Fig. 4.7 Maximum number of users allowed on a network vs. packet length.

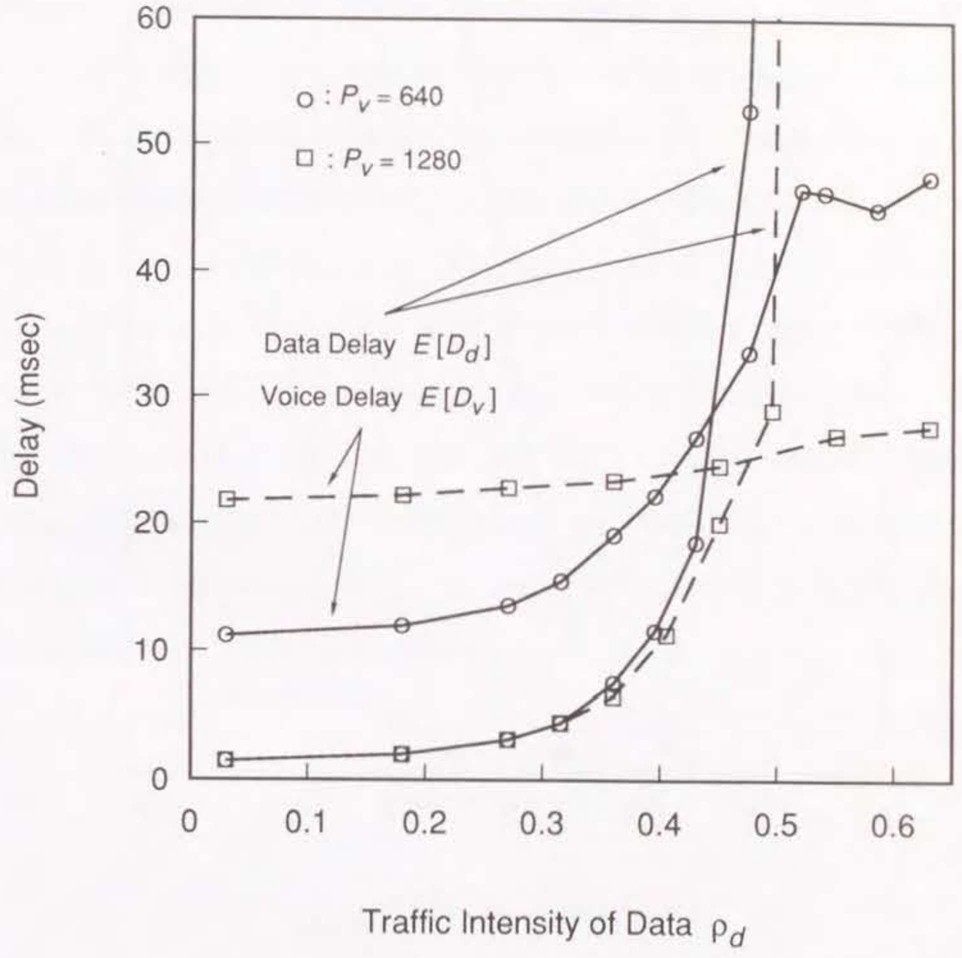


Fig. 4.8 Average transmission delay vs. traffic intensity of data.

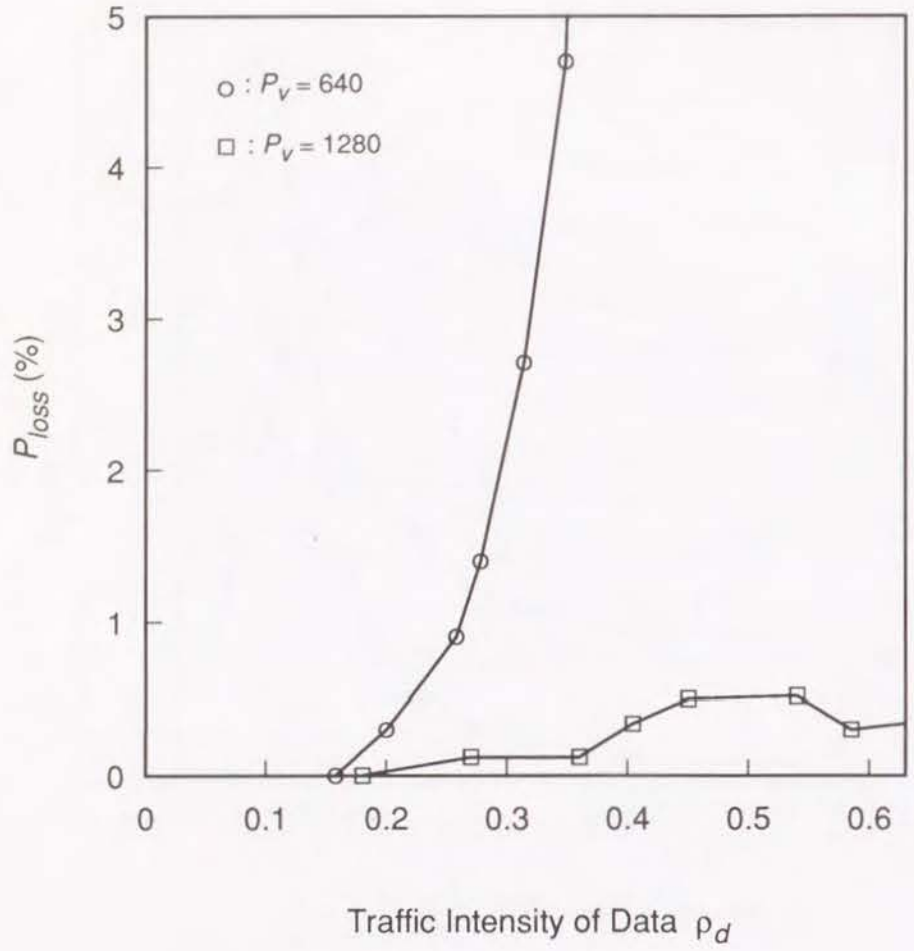


Fig. 4.9 Loss probability vs. traffic intensity of data.

4.6 Conclusions

In this chapter, simulation models have been developed to evaluate performance of packetized voice/data transmission on a token passing ring network. The data users present bursty traffic; voice traffic is modeled as having talkspurts and silences, with generation of voice packets at a constant rate during talkspurts and no packet generation during silence period.

The results suggested that system parameters (e.g., voice packet length, data packet length) can be adjusted to allow token passing ring networks to support both voice and data traffic with acceptable performance, Using the results obtained through simulations, the following issues are examined: selecting the optimum voice packet length, the maximum number of voice calls supported by a network, and the effects of talkspurt/silence lengths and data traffic intensity on network performance.

CHAPTER 5

Approximate Analysis for Bulk Queueing System with Composite Service Discipline

5.1 Introduction

This chapter considers a queueing system where customers arrive in groups according to compound Poisson process and form a single queue. Customers in a group have an identical attribute. Waiting customers are served by a single server in a bulk service fashion according to the predetermined discipline. From the head of the queue, customers with the same attribute are served within a finite bulk size at the same time. Customers having arrived in different groups can be served in the same bulk if they have an identical attribute. On the contrary, customers in a group may not be served simultaneously due to limited bulk size of service. Then, a group may happen to be divided into two or more subgroups. A service is carried out during a time interval between service occasion epochs. The time intervals between service occasion epochs are independently and identically distributed with an arbitrary distribution.

This kind of queueing system can be applied to analyze some of stochastic behaviors in communication system as a mathematical model. Examples are as follows. In a packet switching network, each switching node assembles several packets having the same attribute as a common destination into a frame with fixed length. And gate nodes perform almost the same role to connect multiple local area networks through terrestrial global network or satellite communication network, and in the case of ISDN, all external interfaces will play the same role. The composite service discipline is also applicable to integrated LANs with token-based access methods such as token passing ring, token passing bus and FDDI, in order to make efficient channel use by reducing the overhead part of frames. In those cases, a customer, an

attribute, a server and a service time are regarded as a packet, a destination or a kind of packet, a switching equipment for framing and a time interval of sending a frame, respectively. And a service occasion epoch is regarded as a time epoch that a gate opens, or as a time that a token comes.

The bulk queueing system was first studied by Bailey in [BAIL54]. He considered the single-arrival bulk-service queueing model. Precisely, his model is $M/G^C/1$ in which the maximum number of customers taken for service is constant. He derived the equilibrium distribution of queue length by the embedded Markov chain method. Jaiswal [JAIS60] solved the same model by phase method. The first study of bulk-arrival bulk-service queueing system was done in 1959 by Miller [MILL59]. Since then, many researchers (Bhat [BHAT64], Cohen [COHE69], Neuts [NEUT79], et al.) considered various versions of bulk-arrival bulk-service queueing systems. They treated homogeneous bulk queueing systems (HBQS), that is, they assumed that all customers in the system have an identical attribute. On the contrary, Watanabe et al. [WATA83] introduced Exclusive Group Service (EGS) discipline in $M^X/G^Y/1$ type bulk queueing system, where arriving groups have different attributes though customers in group have an identical attribute.

Our discipline introduced in this chapter is the extended version of EGS discipline[OHTS85]. We consider a composite bulk service discipline which allows different groups to have the same attribute. For a bulk arrival queueing system with a composite bulk service discipline, we approximately analyze the average queue length of customers in the system. In our approach, each class of customers with the same attribute is assumed to form their own queue. There exist as many queues as the number of different attributes. We derive an approximation method for obtaining queue length of each queue just before the service occasion epoch, where probability generating function (p.g.f.) approach and embedded Markov chain method are utilized.

5.2 Model

Let us consider a transportation type bulk-arrival bulk-service queueing system model under composite bulk service discipline.

Customers arrive in groups according to compound Poisson process, that is, the groups of customers arrive according to Poisson process and the sizes of groups, the number of customers in a group, are independent and identically distributed with an arbitrary distribution. Furthermore, all customers of a group have the same attribute. The service, which is in batch of fixed capacity, begins just after a service occasion epoch and the time intervals of service occasion epochs are independently and identically distributed with an arbitrary distribution. The server is able to accommodate as many customers as possible within the capacity if at the back there are the groups of customers whose attributes are the same as the first one in the queue. If there are no customers in the queue at the time of completing the service, next service does not start as soon as the customers arrive at the system but starts after a next service occasion epoch.

In Fig. 5.1, a symbol stands for a customer and the same symbol means an identical attribute. The number figured in a symbol is the group number arriving at the node. Fig. 5.1-(a) shows a situation just before service starting. The queue consists of four groups and the number of customers of the head group is 2. In what follows, we call the customers of the earliest arrival group in the queue as 'head customers' for convenience. Head customers have a claim to be served at first within the bulk size which is 4 in this figure. If the server can serve more customers yet, he accommodates the subsequent customers who have the same attribute as the head customers. Then, the system state changes as Fig. 5.1-(b). In our approach, each class of customers with the same attribute is assumed to form their own queue (see Fig. 5.2). There exist as many queues as the number of different attributes. The service discipline is put into another

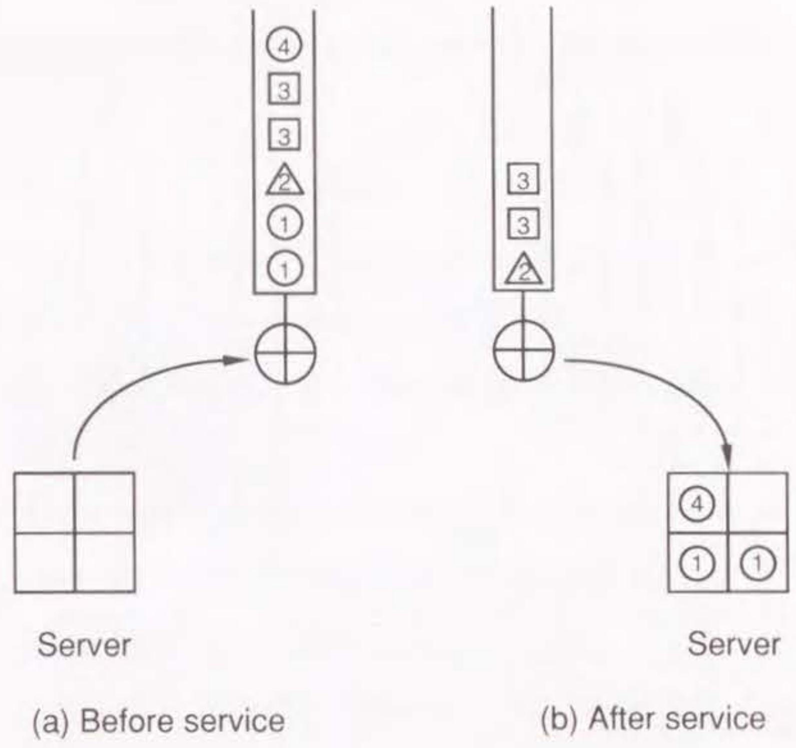
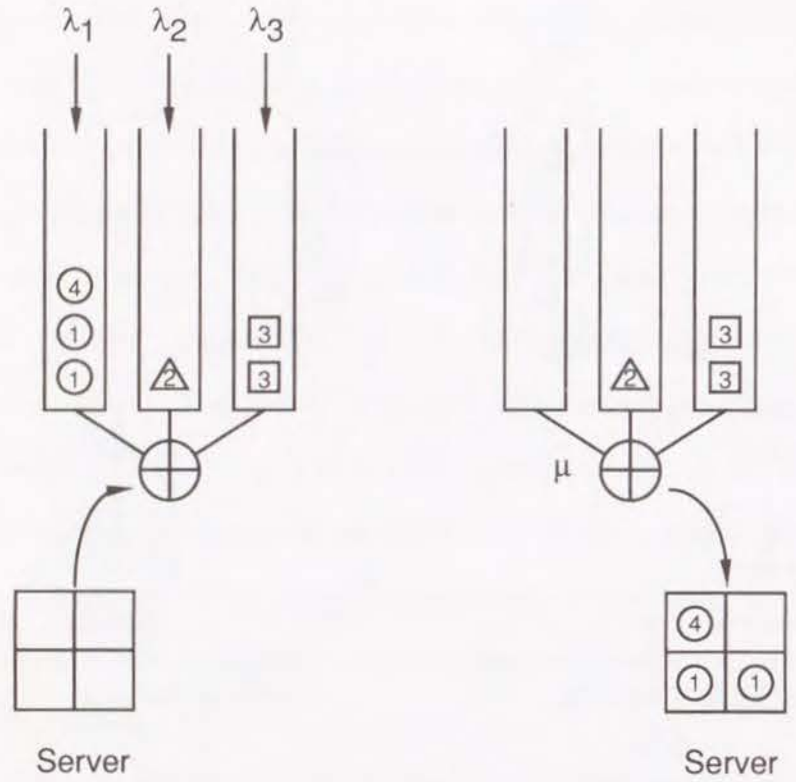


Fig. 5.1 Bulk queueing system model with composite service discipline.



(a) Before service (b) After service

Fig. 5.2 Equivalent model for the bulk queueing system model.

way as follows. Customers in the queue in which the head customers belong to are served within the capacity according to FCFS discipline. Note that FCFS discipline only applies to the customers (or groups) who have the same attribute. For example, in the case of Fig. 5.2, group 4 is served earlier than group 2 and 3, although he arrives later.

Following notations are employed for the parameters of this queueing model.

n : The number of attributes.

λ_i : The arrival rate (Poisson rate) of groups of customers whose attribute is i in queue i ($i = 1,2,\ldots,n$).

λ : The total arrival rate (Poisson rate) of groups of customers ($\lambda = \Sigma \lambda_i$).

b_k^i : The probability that a group of customers whose attribute is i consists of k customers at its arrival, where $k \geqq 1$.

c : The capacity of a server.

$V(t)$: The probability distribution function of the time intervals of service occasion epochs.

5.3 Analysis

5.3.1 Formulation of the queueing model

We propose approximation method of obtaining a probability generating function for the queue length of customers immediately before the beginning of service under a composite bulk service discipline. We pay attention to the behavior of customers who have the same attribute, and then analyze the average length of tagged queue by introducing the probability q that customers in the tagged queue, if any, are served. We assume that queue i is formed by only customers whose attributes are i and that the buffer size of each queue is infinite We introduce the following notations (random variables) to formulate our queueing model.

- V_j : The time interval of the j-th service occasion epoch and $j+1$-st one. we call V_j as j-th time interval.
- $X_j{}^i$: The number of arriving customers at queue i during the time interval V_j.
- $W_j{}^i$: The number of waiting customers in queue i immediately before the j-th service occasion epoch.
- $Y_j{}^i$: The number of customers in queue i who are accommodated during the j-th time interval.
- $B_k{}^i$: The size of k-th arriving customer-group in queue i. Let us call $B_k{}^i$ as the customer-group size in queue i.

In what follows, we pay attention to queue 1 (i.e. customers of attribute 1) and analyze the queue length of queue 1. To make simple, the above random variables are rewritten as, $X_j = X_j{}^1$, $W_j = W_j{}^1$, $Y_j = Y_j{}^1$, and $B_k = B_k{}^1$. Accordingly, the following recurrence relations can be provided for the queue length W_j of customers and the number of served customers Y_j, as shown in Fig. 5.3.

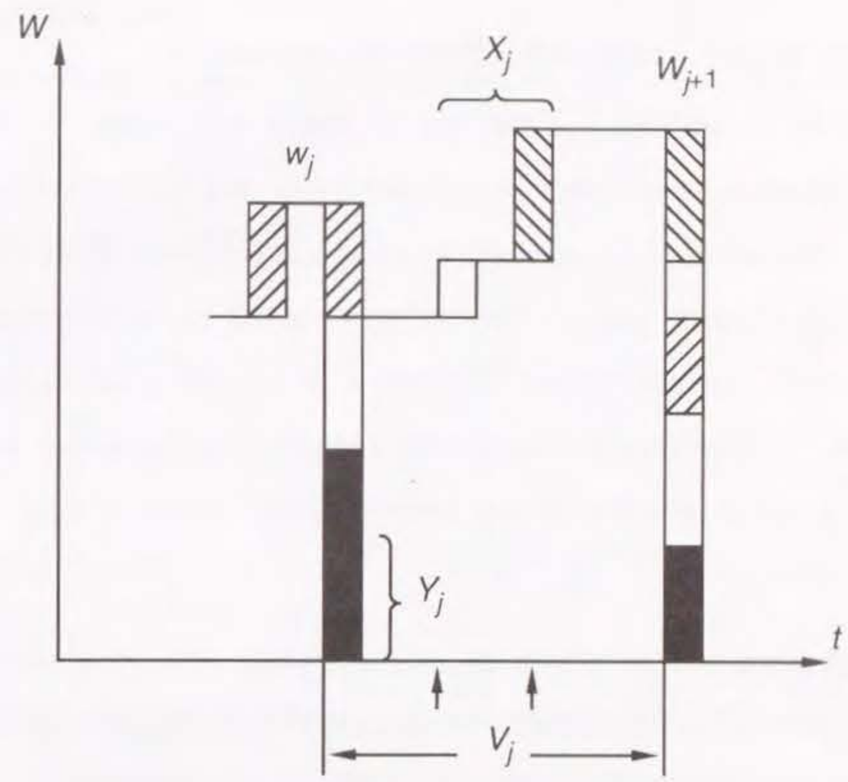


Fig. 5.3 Queue length of customers (W).

$$W_{j+1} = W_j - Y_j + X_j \tag{5.1}$$

$$Y_j = \begin{cases} min[W_j, c] & \text{(the case where customers can be served)} \\ 0 & \text{(otherwise)} . \end{cases} \tag{5.2}$$

5.3.2 Approximate Analysis for Mean Queue Length

Eq. (5.2) shows that the number of served customers, Y_j, depends on whether customers in queue 1 can be served or not rather than the queue length W_j. Therefore, it is necessary to grasp not only the situation of queue 1 but also that of other queue. Precisely, we must know whether the head customers in the system belong to queue 1 or not, in order to analyze this system exactly. However, it is extremely complicated to examine that exactly. Then, we employ the following assumptions in analyzing the model approximately.

Assumption

The probability that the service is available at the j-th service occasion epoch is independent of the queue length W_j, and identical.

It means that, if there are any customers waiting in queue 1 at the the j-th service occasion epoch, $min[W_j, c]$ customers among them is served at this epoch with probability q and the customers in one of the other queues are serviced with probability $1-q$. The probability q is approximately given by using c, b_i (the mean group size of customers in queue i), λ_i and μ (the mean service rate) as below.

$$q = \Big(\sum_{i=2}^{n} \frac{\lambda_i b_i}{c} \Big) / \mu = 1 - \sum_{i=2}^{n} \frac{\lambda_i b_i}{c \mu} . \tag{5.3}$$

This equation is derived under the assumption that a server drops in queue 1,

$$\Big(\sum_{i=2}^{n} \frac{\lambda_i b_i}{c} \Big) / \mu$$

times during μ times service occasions and can collect customers in queue 1.

It what follows, we assume that the system is in steady state. It should be noted that if the subscript j of the random variables is omitted, it means that j need not be considered.

The probability p_k that k customers arrive at queue 1 during a current service is,

$$p_k = Prob[X_j = k] = Prob[X = k]$$

$$= \sum_{k=0}^{\infty} Prob\,[X = k \mid V = t] dV(t) = \int_0^{\infty} \sum_{k=0}^{\infty} exp\,(-\lambda_1 t) \frac{\lambda_1 t}{n!} b_k^{n^*} dV(t) \tag{5.4}$$

where

$$b_k^{n^*} = Prob\,[B_1 + B_2 + \ldots. + B_n = k] . \tag{5.5}$$

The probability generating function, p.g.f., $P(z)$ of X is derived from both the p.g.f. $B(z)$ of B, i.e., the size of a customer group at its arrival time, and $V(t)$.

$$P(z) = \sum_{k=0}^{\infty} p_k z^k = \int_0^{\infty} exp\,(-\lambda_1 t\,(1 - B(z)) \, dV(t) \ . \tag{5.6}$$

$$B(z) = \sum_{k=0}^{\infty} b_k z^k \qquad (|z| \leqq 1) \ . \tag{5.7}$$

According to Eqs. (5.1) and (5.4), the following equation is held.

$$Prob[W_{j+1} = k \ \ W_j = n, Y_j = m] = Prob[X_j = k + m - n] = p_{k+m-n} \quad (m \leqq n). \tag{5.8}$$

Multiplying by z^k on both sides of the above equation and adding over k, we have,

$$\sum_{k=0}^{\infty} Prob\,[W_{j+1} = k \mid W_j = n, \ Y_j = m] z^k = P(z) z^{n-m} \tag{5.9}$$

Furthermore, the following equation is obtained, if W_j and Y_j are removed.

$$\sum_{k=0}^{\infty} Prob\,[W_{j+1}=k]\,z^k =$$

$$= P(z) \sum_{k=0}^{\infty} Prob\,[\,W_j = n,\ Y_j = min\,(n,c)]\,z^{n-min(n,c)} + Prob\,[\,W_j = n,\ Y_j = 0]\,z^n$$

$$= P(z) \sum_{k=0}^{\infty} Prob\,[\,W_j = n,\ SV(j)]\,z^{n-min(n,c)} + Prob\,[\,W_j = n,\ NSV(j)]\,z^n\ . \tag{5.10}$$

Where, $SV(j)$ and $NSV(j)$ denote the event that the customers in queue 1 can be and can not be served during the j-th time interval, respectively. Then, we have

$$Prob[W_j = n, SV(j)] = q\,Prob[W_j = n]\ . \tag{5.11}$$

$$Prob[W_j = n, NSV(j)] = (1-q\,)\,Prob[W_j = n]\ . \tag{5.12}$$

By introducing eqs. (5.11) and (5.12), eq. (5.10) is rewritten as follows.

$$\sum_{k=0}^{\infty} Prob\,[W_{j+1}=k]\,z^k$$

$$= P(z) \sum_{k=0}^{\infty} (q\,Prob\,[\,W_j = n]\,z^{n-min(n,c)} + (1-q)\,Prob\,[\,W_j = n]\,z^n\,)$$

$$= P(z) \left(q \sum_{k=0}^{c-1} Prob\,[\,W_j = n] + q \sum_{k=c}^{\infty} Prob\,[\,W_j = n]\,z^{n-c} + (1-q)\,Prob\,[\,W_j = n]\,z^n \right) \tag{5.13}$$

Defining $w_k = lim(j \to \infty) Prob[W_j = k]$ from the steady state assumption, the p.g.f. $R(z)$ is given as follows.

$$R(z) = \sum_{n=0}^{\infty} w_n z^n = A(z)\,/\,D\,(z) \tag{5.14}$$

where

$$A(z) = q \sum_{n=0}^{c-1} w_n\,(z^c - z^n\,) \tag{5.15}$$

and

$$D(z) = z^c\,(P^{-1}(z) - (1-q)) - q\ . \tag{5.16}$$

The unknown variables w_n ($n=0, 1, \ldots, c-1$) can be determined from the facts that $R(z)$ is analytical in $|\,z\,| \leqq 1$ and that $D(z)=0$ has the c zeroes in $z \leqq 1$ by Rouche's theorem. If we assume that $D(z)=0$ has 1 and ξ_i ($i=2, 3, \ldots, c$) as its simple zeroes, w_n ($n=0, 1, \ldots, c-1$) are obtained as the solution of the following linear equations. Note that, in what follows, the i-th derivative of p.g.f.'s are denoted like as $P^{(i)}(z)$.

$$A^{(1)}(1) = D^{(1)}(1)\ . \tag{5.17}$$

$$\sum_{n=0}^{c-1} w_n\,(\xi_i^{\,c} - \xi_i^{\,n}\,) = 0 \qquad (i=2, 3, \ldots, c)\ . \tag{5.18}$$

Eq. (5.17) is derived by the equation $lim(z \to 1)\,R(z) = 1$.

Then, the average queue length of customers in queue 1, $E[L_1]$ is obtained as follows.

$$E[L_1] = R^{(1)}(1) = \frac{A^{(2)}(1)D^{(1)}(1) - A^{(1)}(1)D^{(2)}(1)}{2\,(D^{(1)}(1))^2}\ . \tag{5.19}$$

where,

$$A^{(1)}(1) = q \sum_{n=0}^{c-1} w_n\,(c-n\,) \tag{5.20}$$

$$A^{(2)}(1) = q \sum_{n=0}^{c-1} w_n((c-1\,) - n(n-1)) \tag{5.21}$$

$$D^{(1)}(1) = c\,q - P^{(1)}(1)\ , \tag{5.22}$$

and

$$D^{(2)}(1) = c(c-1)q - 2cP^{(1)}(1) + 2(P^{(1)}(1))^2 - P^{(1)}(1). \tag{5.23}$$

5.4 Numerical Examples and Evaluations

We now evaluate the approximation method by comparing them with simulations. Table 5.1 shows the validity of our simulation experiment, where simulation results are compared with exact theoretical values in the case that the number of attributes is equal to one (i.e., HBQS case), and that the time intervals of service occasion epochs are constantly distributed.

Following numerical results are derived in the case that the sizes B^i of customer groups at their arrival time are according to a geometric distribution with the mean equal to b_i $(=1/p)$, i.e., $b_k^i = Prob[B^i = k] = p(1-p)^{(k-1)}$ $(k=1, 2, \ldots)$.

At first, we show the numerical results in the case where the time intervals of service occasion epochs are according to an exponential distribution with parameter μ and that the number of attributes is equal to two. Fig. 5.4 shows the relationship between the average queue length of customers in queue i, $E[L_i]$ $(i=1,2)$ and the Poisson arrival rate of groups of customers whose attribute is 1 λ_1, in the case where $\lambda_2=0.1$, $b_1=b_2=3.0$, μ = 1.0 and c = 5. Lines and symbols present analytical results and simulation results respectively. From this figure, we find that the theoretical results are well verified by the simulation results. Further it is found that $E[L_1]$ is affected by the change of λ_1, but that $E[L_2]$ is not so much. Namely, in such a case, the mean queue length of one queue is not much influenced by the arrival rate of group-customers in other queues.

Fig. 5.5 shows the relation between $E[L_1]$ and c in the case where μ= 1.0, λ_1 = λ_2= 0.1 and $b_1=b_2$= 5.0. As illustrated in this figure, it is recognized that if c is sufficiently larger than b_1 $(=b_2)$, $E[L_1]$ changes little. That is because all waiting customers in the same queue are considered to be served at once.

Next, we consider the case where the service occasion epochs have constant time intervals and the number of attributes is three. In Fig. 5.6 the relationship between $E[L_1]$ and λ_1 are shown with $\lambda_1=\lambda_2=\lambda_3$, $c=5$ and $\mu=1.0$ under various values of b_1 $(=b_2=b_3)$. It is confirmed that slight increase of the bulk size of arriving customers causes the large queue length in the case of heavy traffic. Furthermore, even if the mean arrival number per time (i.e. λ_i b_i) is the same, the larger the bulk size becomes, the more the queue length increases.

Fig. 5.7 shows the mean queue length of each queue as a factor of b_1. Similar to Fig. 5.4, we can say that the mean queue length of one queue is not so much influenced by the mean customer-group size in other queues.

Fig. 5.8 shows the mean queue length for various λ_1 in the case where $\lambda_1 b_1$ is fixed to 0.4, $b_2=b_3=$ 4.0, $c=3$, $\mu=1.0$. The fitness of our approximation is not good for the case when the average bulk size of another attribute is too large.

In general, our approximation results are well verified by the simulation results shown in these figures but have tendency to underestimate the simulation ones, which is caused by the approximation of the probability q_i. In eq. (5.23), we assumed $\lambda_i b_i /c$ as the mean frequency of service per time for queue i, but this value is smaller than exact one because a server does not always serve c customers. Therefore, the value q of our method is greater than the real one and then mean queue length is underestimated.

Many examples show our approach provides approximation results with rather good accuracy, and, in general, has the similar tendency with these examples.

Table 5.1 Comparison between exact and simulations for HBQS model.

($\mu = 1.0$, $b = 3.0$, $c = 3$)

λ	$E[L]$	
	Exact	Simulation
0.1	0.45	0.45
0.2	0.98	0.98
0.3	1.60	1.61
0.4	2.37	2.40
0.5	3.38	3.33
0.6	4.80	4.78
0.7	7.06	7.03
0.8	11.41	11.39
0.9	24.13	24.27

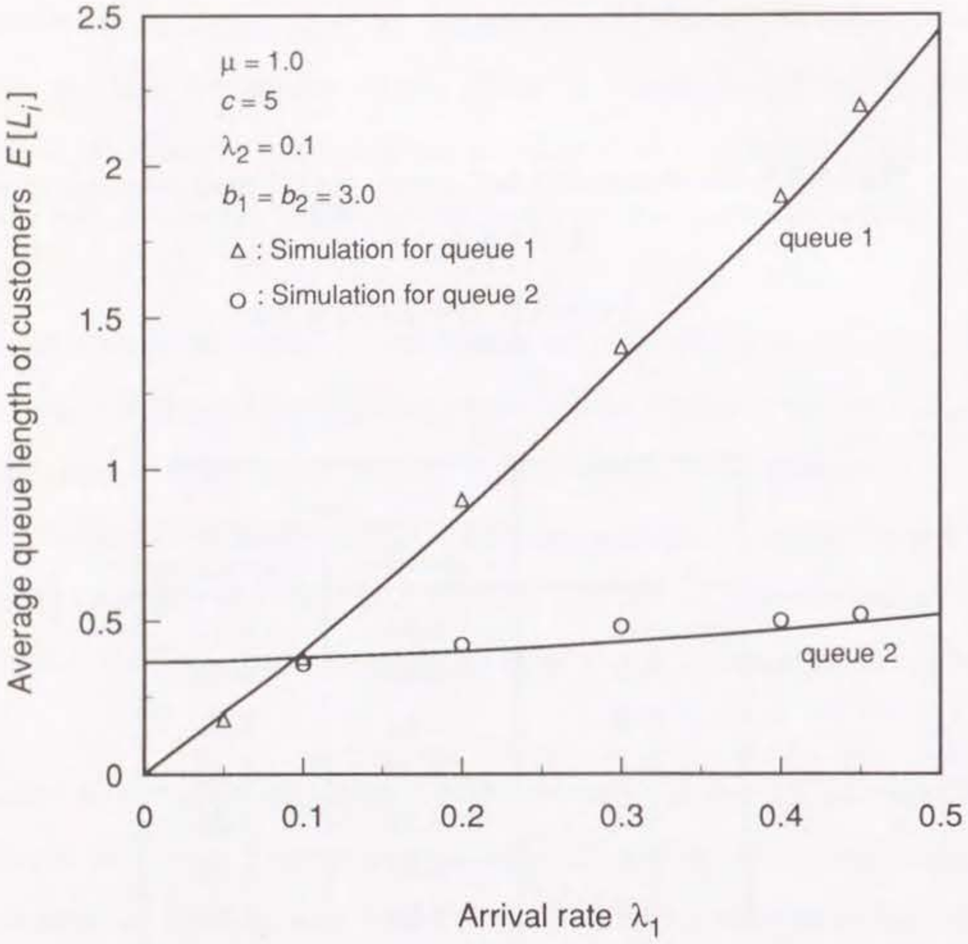


Fig. 5.4 Average queue length vs. arrival rate λ_1.

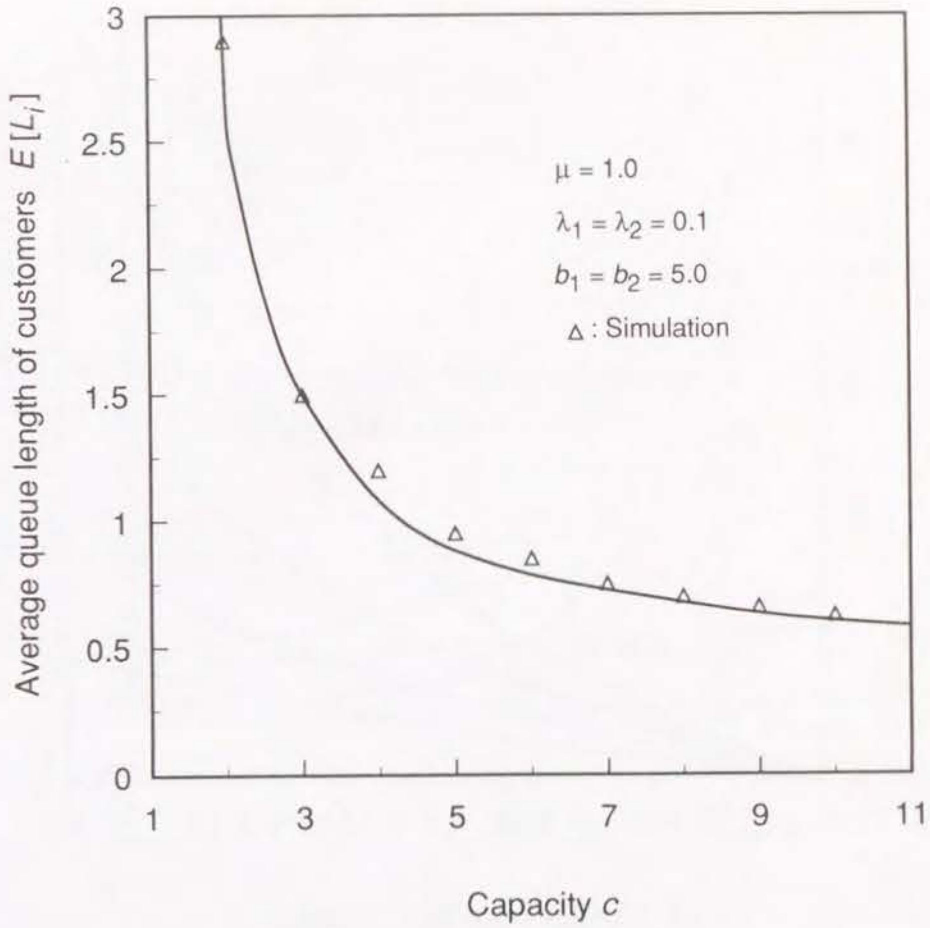


Fig. 5.5 Average queue length vs. capacity of a server.

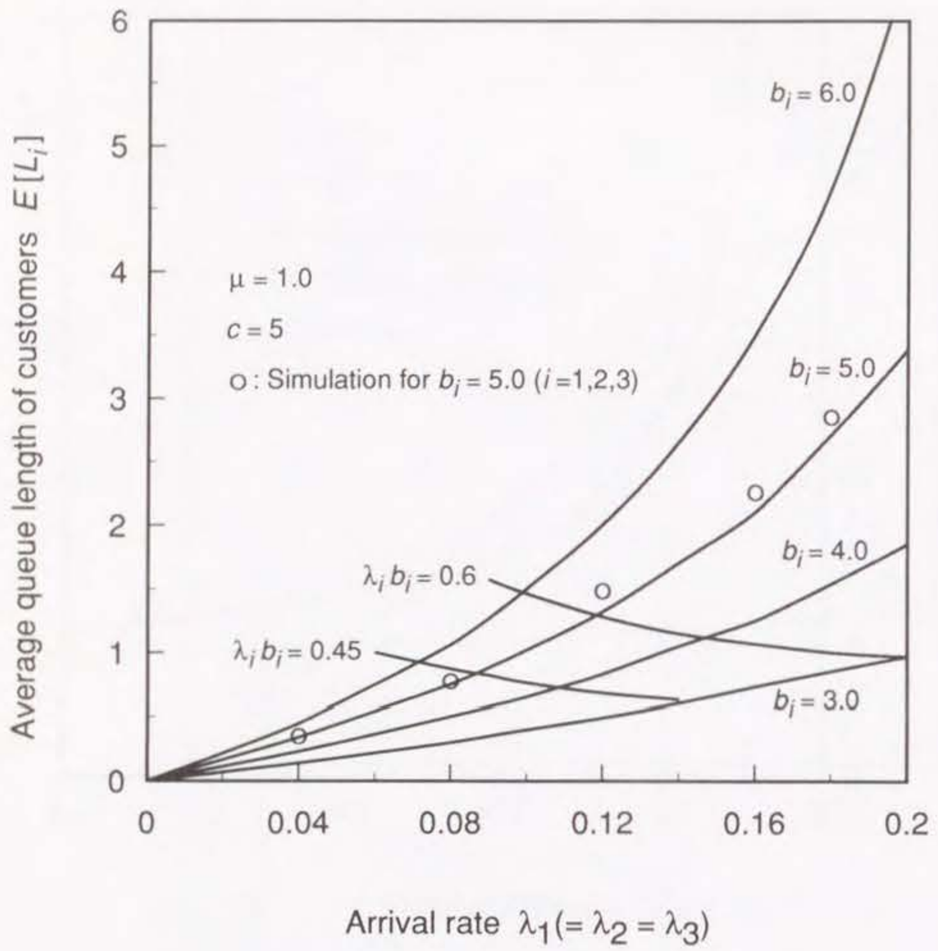


Fig. 5.6 Average queue length vs. arrival rate
: symmetric arrival model.

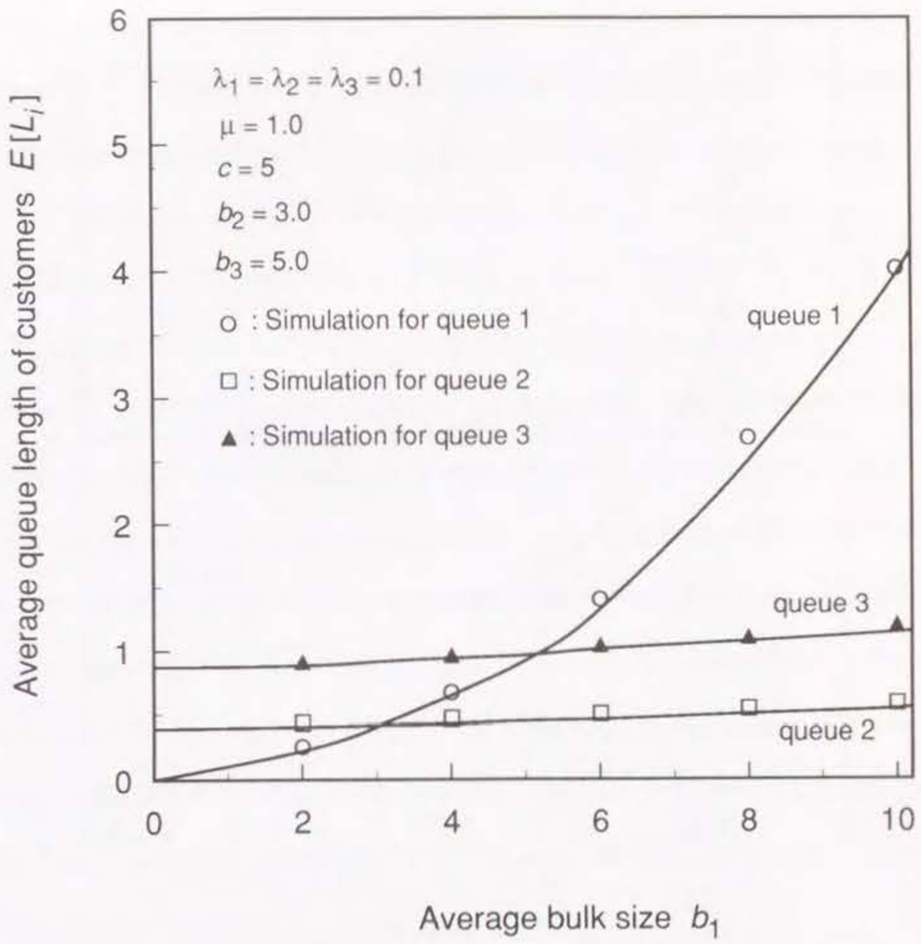


Fig. 5.7 Average queue length vs. average bulk size of attribute 1, b_1.

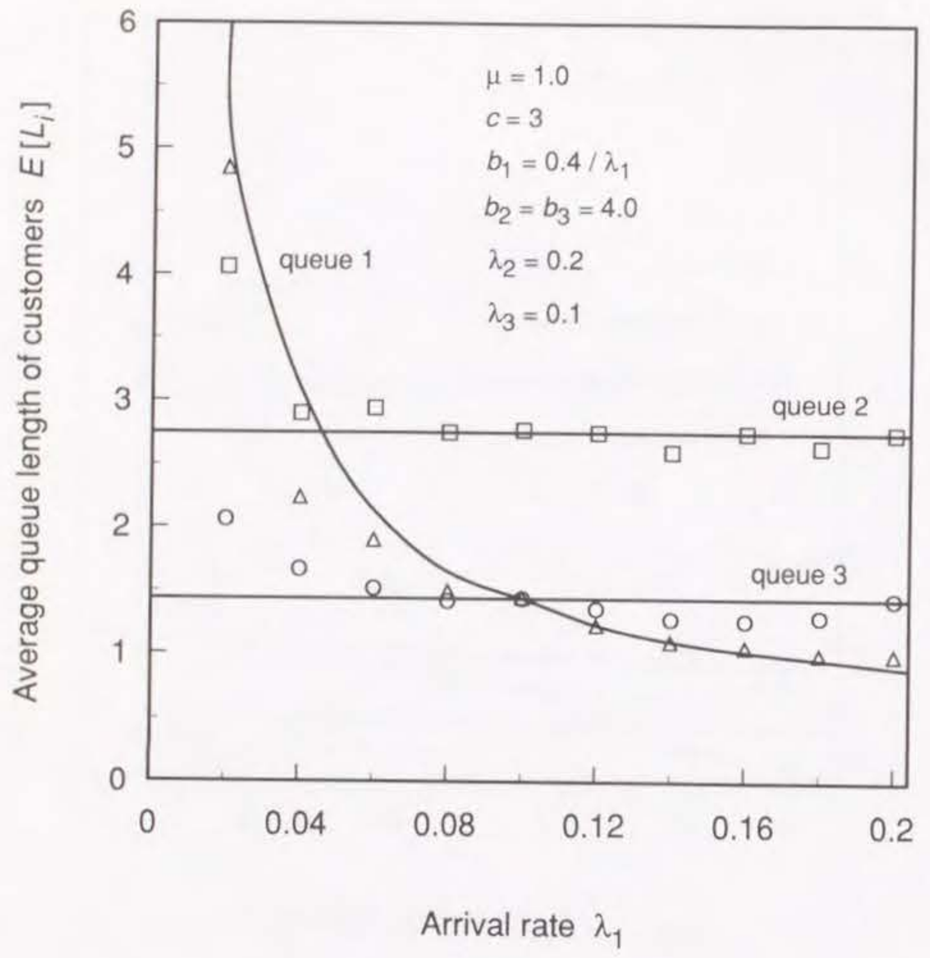


Fig. 5.8 Average queue length vs. arrival rate λ_1, where b_1 is inversely proportional to λ_1.

5.5 Conclusion

In this chapter, we have considered a transportation type bulk-arrival bulk-service queueing system with composite bulk service discipline as a fundamental research on future integrated communication networks. In this system, customers arrive in groups and customers in a group have an identical attribute. From the head of the queue, customers with the same attribute are served within a finite bulk size at the same time.

This kind of queueing system can be applied to analyze some of stochastic behaviors in communication system as a mathematical model. In a packet switching network, each switching node assembles several packets having the same attribute as a common destination into a frame with fixed length. And gate nodes perform almost the same role to connect multiple local area networks through terrestrial global network or satellite communication network, and in the case of ISDN, all external interfaces will play the same role. The composite service discipline is also applicable to an integrated token passing ring network in order to make efficient channel use by reducing the overhead part of frames.

For such a bulk service queueing system, the average queue length of customers in the system was approximately analyzed. In our approach, each class of customers with the same attribute is assumed to form their own queue. Our approximation method derived the queue length of each queue just before the service occasion epoch, where probability generating function approach and embedded Markov chain method were utilized. Many examples showed that our proposed method provides high approximation accuracy by comparing them with the results of simulations. With the proposed method, it has become possible to evaluate the characteristic quantities of the system considered.

CHAPTER 6

A Packet Switch Architecture with Channel Group Virtual Circuit Scheme in High-Speed Communication Networks

6.1 Introduction

Advances in fiber optic technology and hardware based fast-packet switching techniques are quickly making the transport of high-bandwidth application such as video services a reality. However, the current need to perform the switching and packet buffering using electronics imposes an important constraint on the maximum rate at which data can be transmitted over a single channel. It is possible, however, to still tap the tremendous bandwidth of a single transmission fiber by simultaneously multiplexing multiple logical channels over a single physical channel using a technique such as wave division multiplexing.

In this chapter, we thus consider the case in which multiple parallel channels, each operating at a speed of 1-100 Mbps, connect a pair of adjacent switching nodes. There are several reasons to consider such a switch [CHLA88]. First, as discussed above, such techniques can more fully exploit the tremendous potential bandwidth of fiber optics. Second, switching a larger number of slower speed channels, obviates the need for ultra-fast electronic technologies; high switching capacities are obtained by essentially duplicating lower speed channels in less expensive technology rather than implementing single channel switches in more expensive technology. Finally, multichannel architectures have the advantage of providing higher reliability by offering alternative virtual channels.

A major concern for high-speed switch architectures is packet congestion in the switching node. Currently proposed switching architectures such as the

store-and-forward type, Banyan type [HUI87], [TURN86] and Bus-matrix type [NOJI87], [YEH87] are based on a "single channel allocation" policy when a switch transmits packets to another switching node. In such a single channel routing policy, the routing information contained in a packet determines a unique, predetermined channel among several channels between two switching nodes. In such networks, as illustrated in Fig. 6.1, congestions may take place at an output port (due to a temporary traffic load exceeding the channel capacity), while other channels between the same two switching nodes are underutilized. This congestion is inevitable in any switch architecture when the traffic is bursty. The problem is particularly acute in single channel architectures, however, since there is no option of transmitting congested traffic over an alternate, less congested, channel. Thus, a new channel allocation protocol for the control of congestion problem in a multichannel system is required to provide an efficient channel use and low switching delay.

One way to solve this congestion problem is to allow a packet to switch from a congested transmission channel to an uncongested one as shown in Fig. 6.2. Pattavina [PATT88] introduced a two step bandwidth allocation scheme in which a set of channels between two adjacent switching nodes is viewed as a single virtual channel by routing entities. At connection set up time, packet routing is determined on a virtual channel basis, and at transmission time, a packet selects an appropriate outgoing channel from among those in an available channel group. This scheme not only solves the congestion but also makes an efficient use of transmission channels. Pattavina presents a switch architecture implementing the scheme in Batcher-banyan switch, but the switch has following disadvantages. One is that each channel has a different priority level for channel selection. If many of packets are transmitted on the higher prioritized channels, the traffic rate of each outgoing channel and incoming channel can be unbalanced significantly. The second is that a stream of packets which have the same source-destination pair may be assigned to

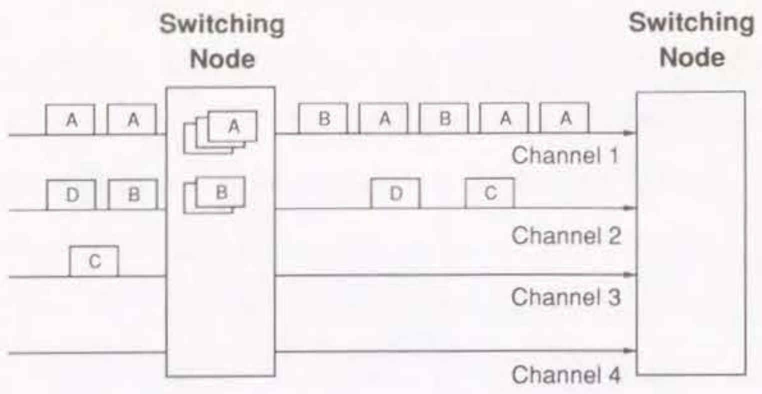


Fig. 6.1 Packet congestion in a switch for single channel allocation policy.

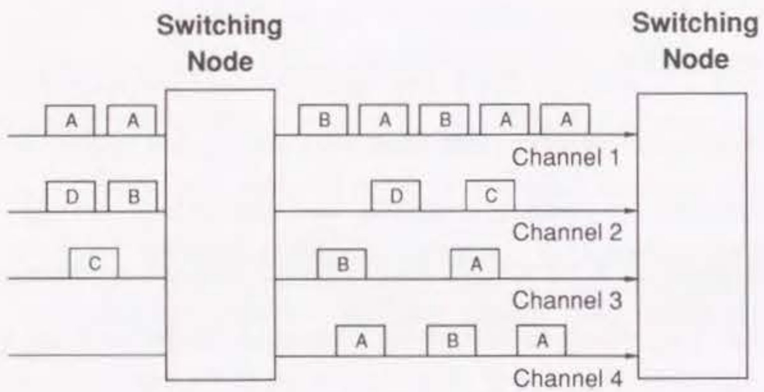


Fig. 6.2 Dynamic channel allocation.

different channels within a channel. Thus, the issue of packet sequencing must be addressed as a strict problem to be solved.

In this chapter, we introduce a new channel allocation scheme called the channel group virtual circuit (CG-VC) scheme which is based on a two step channel allocation[OHTS90, 91a]. The scheme not only makes an efficient use of the transmission channels, but also keep the stream of packets being transmitted on the same channel in order if the channel is not congested. We propose a new switch architecture to realize the CG-VC scheme. The switch uses a revised ring protocol for a high speed LAN in which a packet is switched to one of the appropriate output ports via optical switching rings in the node.

As we will see, the multi-ring switch distributes the traffic over the multiple logical outgoing channels of the same channel group in such a way as to avoid currently congested logical channels; it thus reduces transmission queueing delay (via its load balancing effect) and also decreases packet loss due to buffer overflow in times of congestion. We present a simple queueing model for preliminary examination for the performance of the switch. We approximately analyze transmission delay, and the results are compared with simulation results.

The remainder of this chapter is as follows. In the next section, we investigate a control scheme for multichannel allocation. In section 6.3, we describe the architecture of the multi-ring switch. A performance model of the switch is presented and is evaluated in section 6.4. Finally, in section 6.5, conclusions are drawn.

6.2 Channel Group Virtual Circuit Scheme

Conceptually, the best way to make use of multiple channels in a multichannel system is to have a very high-speed processor controlling all packets in the switch and have this processor allocate packets to channels. However, given current technology, it is impossible for a processor to operate at the rate required when multiple gigabit rate channels are aggregated in a switch and packet sizes are relatively short (e.g., an ATM cell is 53 octets long). An appropriate goal, then, is to provide the load balancing capability in hardware, using a simple form of hardware load balancing if at all possible. The CG-VC architecture provides such hardware-based load balancing.

The major problem with two step channel allocation is that the sequence of packets which have the same source-destination pair may delivered out of order on an end-to-end basis. It is not yet clear that the packet sequence integrity is a critical requirement for high speed networks, and we allow here for packets being out of order. In this case, packet resequencing must be employed by the higher layer at the destination terminal. Even though, we believe that it is better to maintain the packet sequencing as much as possible.

Given the above considerations, we now introduce our CG-VC scheme. Fig. 6.3 shows the basis of this scheme. The switch consists of several switching modules. Parallel logical channels connecting a pair of adjacent switching nodes are carried over one or more physical channels connecting the two nodes. In the figure, there are 2 outgoing channels groups (CGs) each consisting of 3 channels. When a call occurs, a virtual circuit is set up by deciding a outgoing transmission channel group.

At transmission time, packets which arrive from incoming channels of group *a* normally use the outgoing channels connecting to switch module SM-A. For example, a packet whose address is identified as CG1 is allocated to the corresponding output channel in SM-A if the channel is not congested. When

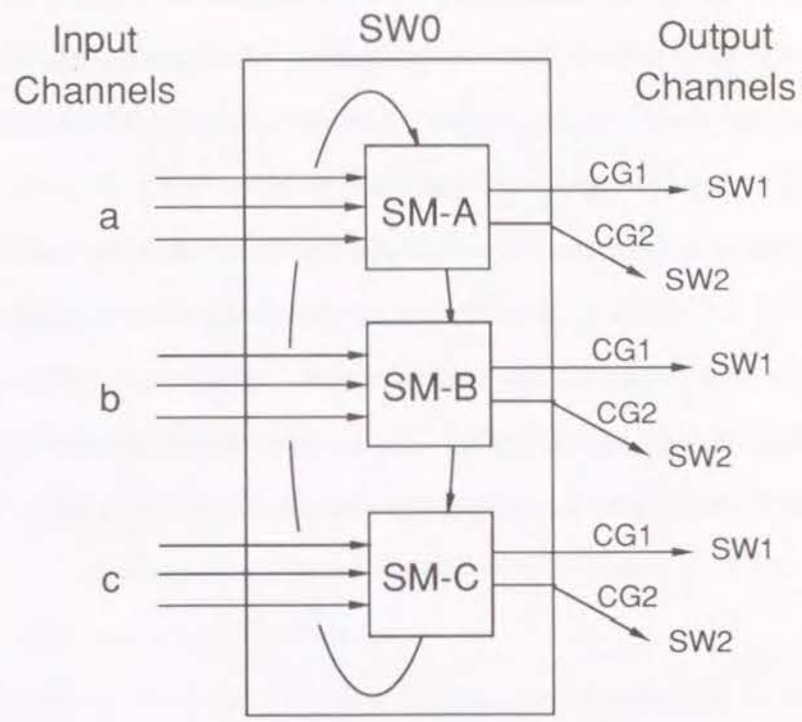


Fig. 6.3 Channel group virtual circuit (CG-VC) scheme.

the target channel is congested, however, the packet is forwarded to next switching module. The packet moves around the switching modules until it is allocated a non-congested channel of CG1. In this scheme, an incoming packet behaves as if it has prioritized multiple transmission channels, that is, a packet arriving at an incoming channel searches for an appropriate output channel in SM-A, SM-B, SM-C in order. We call the policy of determining the outgoing physical channel as the channel group virtual circuit (CG-VC) scheme.

In the light traffic case, all packets will be switched to a highest prioritized output channel; packets having the same source-destination pair (i.e. the same VC) will thus be transmitted on the same channel and will arrive at the destination in order. Only in the case of congestion, will a packet be routed away from its highest priority output channel. In the next section we discuss a switch architecture to implement this load balancing scheme.

6.3 Switch Architecture

6.3.1 Switch Structure

The structure of the multi-ring switch is shown in Fig. 6.4. The switch consists of input ports, output ports, overflow ports and optical fiber rings. An incoming/outgoing channel is connected to a corresponding input/output port. That is, packets from a given channel arrive at an identified input port. An input port is connected to one of the rings, and it simply forwards incoming packets onto the ring. An output port receives a packet from the ring and forwards it to an outgoing channel.

An overflow port is also positioned between rings, which allows for one-way transmission between ring modules. Every ring has two overflow ports, one for sending packets to another ring and the other for receiving packets from another ring. We call the former an output overflow port and the later an input overflow port for a given ring module.

As shown in Fig. 6.4, the switch is constructed by superposing ring modules, each consists of an optical ring, and input, output and overflow ports being connected to the ring. By adding new modules, the switch can modularly grow to provide a larger switching capacity.

6.3.2 Packet Address

Since the proposed switch architecture is based on group channel selection, the routing information is different from that in conventional switching systems. When a packet enters a switch a routing address is attached to the packet identifying the output channel group to be transmitted. This means that the routing information does not prescribe a specific output channel itself but rather only a specific channel group. The packet packet will be transmitted in one of the channels associated with the identified channel group, as discussed below.

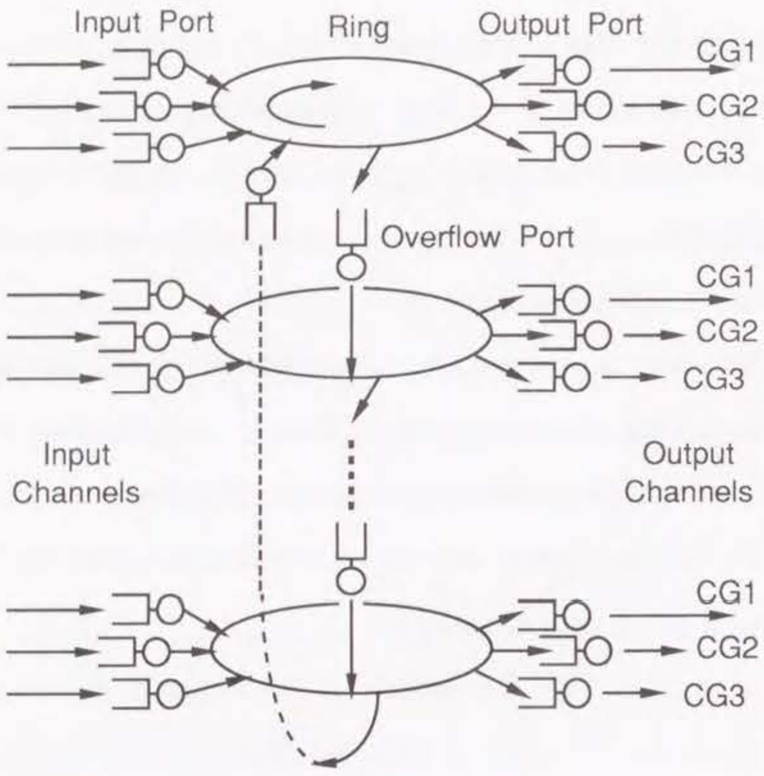


Fig. 6.4 Switch structure.

6.3.3 Routing Mechanism

The traffic control mechanism (i.e., the ability to rout around congested output ports) proposed in this chapter is performed via the routing mechanisms of a multi-ring switch. The mechanism operates as follows.

A packet first arrives at an input port and is stored in its buffer. An input port forwards the stored packet onto a ring in first-in-first-out fashion according to the ring transfer protocol. We will briefly consider this protocol later.

Each output port has a small buffer. The output port, while repeating the incoming signal stream, checks the ring for packets which it should receive. If the routing address of the packet matches the port's individual address and if the buffer of the port has enough space to receive the packet, the port receives the packet into its buffer and eventually transmits the packet. If the port does not have enough buffer space or if the packet was not destined to the port, the packet passes undisturbed past the port on the ring.

An output overflow port sweeps the ring by receiving any packet which is not received by any output port on the ring. The output overflow port then forwards the packets to the next ring as an input overflow port of the next ring module. The size of the overflow port buffer depends on the ring transfer protocol and the differences in transmission capacity between output channels and an optical ring. For example, if the ring's transfer protocol gives the overflow port an absolutely higher priority than input ports on that ring, only one output overflow port buffer is required, since the arrivals of overflowed packets are less than ring capacity.

6.3.4 Ring Protocol

In this chapter we will not examine the ring access protocol in detail since it is outside the scope of this chapter (we focus here in this chapter on the CG-

VC concept, not the ring access protocol). However, we believe that conventional ring protocols for optical LANs would be easily adaptable for the proposed switch; indeed, ring-based switching has already been proposed and implemented by others as the underlying switching fabric for high-speed switching. We also note that an error detection and retransmission mechanism may not be suitable for high speed transmission, and thus a destination removable slotted ring protocol (e.g. Orwell ring [FALC85]) or revised versions of it would be suitable for the switch.

6.3.5 Network Switch Design

In some systems, each channel group may not have an equal number of parallel channels. To adapt the multi-ring switch to such systems, two approaches towards switch design can be taken. One approach is to allow the ring to have two output channels for a certain channel group attached to the same ring module. Similarly, each output channel group need not be connected to all ring modules. For example, if two identical group ports are attached to a ring, packets with the identical addresses will be transmitted via upper stream channel in the case of light traffic and the second channel will be used only in the case of heavy traffic (i.e., the case in which the upstream port can not handle all incoming packets.) Similarly, if a packet has no appropriate output ports on a ring, it is simply passed to the next ring via an overflow port.

Another approach is to build a network switch using a two stage multi-ring switch in which the second stage switch operates to reduce the number of output ports; we call this switch a concentrator switch. A concentrator switch has the same structure as that of the first stage switch except for having only one output channel group. The first stage switch is connected to several concentrator switches, and forwards a packet to an appropriate concentrator switch according to a routing address of the packet.

6.3.6 Advantages of the Switch

We believe the proposed switch has the following advantages.

1) The switch provides efficient channel utilization and a high speed packet delivery by distributing packets among all available transmission channels.
2) Since the switch is constructed by a number of ring modules, it is easier to build a flexibly sized switch; the switch can grow larger incrementally by adding new modules.
3) By implementing the ring module in hardware using registers, the switch can support high speed transmission. Moreover, if the module provides bit parallel transmission by using multiplexed, bit serial rings, each module can handle the very high speed packet delivery of Gbit/s.
4) Since the N-input N-output switch consists of N input switch interfaces, N output switch interfaces, M ($M<N$) optical rings and M ring-to-ring interfaces, complexity of the switch grows only linearly with N.
5) In the case that an output port does not receive packets due to a failure in the output channel, packets are routed to another appropriate channel automatically.
6) In single channel allocation systems, the output port address for an arriving packet is typically determined by the routing controller prior to the packet entering the switch fabric. The choice of an outgoing channel is rather complex from the point of view of maintaining an optimal load distribution for a channel group. On the contrary, the proposed switch does not require such a beforehand channel selection.
7) The switch has a self-routing mechanism.

8) The switch can handle asynchronous packets with variable length since the switching protocol is based on the ring protocol for multiple access LANs.

6.4 Performance Evaluation

6.4.1 Performance Model

In this section, we present a simple queueing model for preliminary examination of the performance for the proposed switch. In the performance model we consider the case that the number of the output virtual link is 1, the number of the outgoing channels is N, an each channel is connected to its own ring (i.e. number of rings is N). We assume that the buffer size of input buffer and overflow buffer are infinite, but that the size of the output buffers restricted to K. Note that it is the (purposefully) restricted output buffer size that permits hardware-based load balancing and thus K is as key parameter in determining the performance of the switch .

Fig. 6.5 illustrates the queueing model. In the model, we consider each ring and output channel as a single server, and all input ports and overflow port on the same ring are gathered to a single queue of the ring. In this case, packets arrive to i-th ring switch from incoming channels at rate λ_i. The arrival process of incoming packets is assumed to be a Poisson process, and packets are transmitted to the ring under an FCFS discipline. The transmission time on the ring has an exponential distribution with mean $1/\mu_R$. If a transmitted packet finds K packets in the output buffer, it overflows onto the next ring module and is stored immediately in the input buffer of the next ring. The arrival rate of the overflowed packets at ring i is denoted as λ'_i, and the offered traffic load of ring i is denoted as G_i, which is the sum of λ_i and λ'_i. The packets accepted by the output ports are transmitted to the outgoing channel. The transmission time of the outgoing channel (i.e., service time of output server) obeys the exponential distribution with mean $1/\mu_C$.

6.4.2 Approximation Analysis

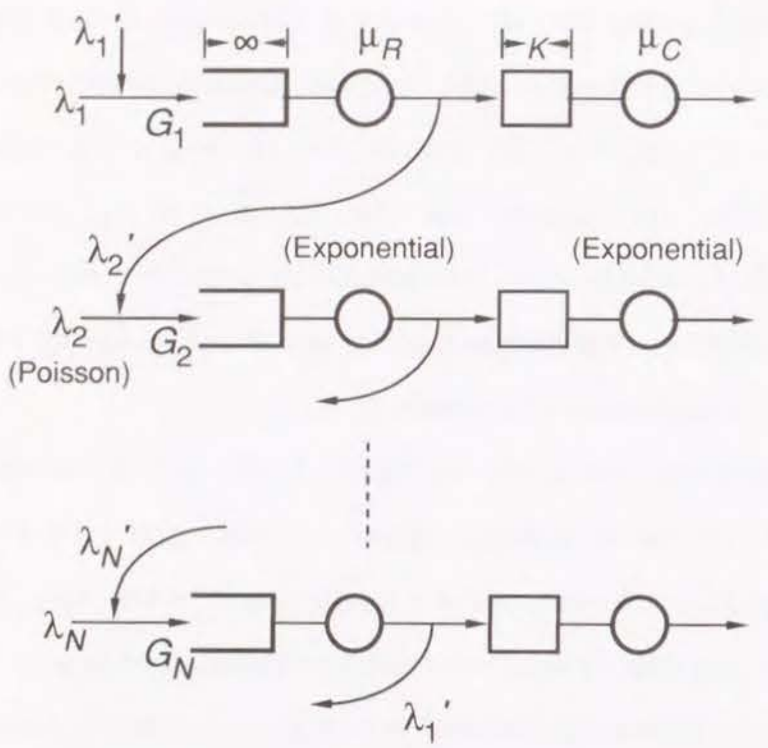


Fig. 6.5 Queueing model.

In the above model, arrivals of overflowed packets are not strictly a Poisson process. However, if we assume the process to be Poisson, the pair of a ring and output channel system together become a tandem queueing system with loss, M/M/1 /∞ and M/M/1/K, and we obtain closed form equations.

The relations between G_i s ($i=1, \dots, N$) is as follows.

$$G_i = \begin{cases} \lambda_i + \lambda'_{i-1} = \lambda_i + G_{i-1} Prob(\text{A packet overflows at ring } i-1) & (\text{for } i=2,\dots N) \\ \\ \lambda_1 + \lambda'_N = \lambda_1 + G_N Prob(\text{A packet overflows at ring } N) . \end{cases} \tag{6.1}$$

The probability that a packet overflows at ring i is equivalent to the the probability that there are K packets in the output buffer i-1, $P_{i\text{-}1,K}$, and is given using M/M/1/K formula as,

$$P_{i\text{-}1,K} = (1-\rho_{i\text{-}1})\, \rho_{i\text{-}1}{}^{K} / (1-\rho_{i\text{-}1}{}^{K+1}) \tag{6.2}$$

where,

$$\rho_{i\text{-}1} = G_{i\text{-}1} / \mu_C . \tag{6.3}$$

From equations (6.1)-(6.3), we derive the polynomial equations in G_i of the order $(K+1)^N$. By solving this equation, G_i is numerically obtained.

We define the switching delay as the time interval from the time packet arrives at the switch to the time at which it is transmitted on an outgoing channel. An average switching delay for packets corresponding to λ_i is denoted as D_i. For example, D_1 is given as

$$\begin{aligned} D_1 &= (1-P_{1,K})\,(D_{R1}+D_{C1}) + P_{1,K}(1-P_{2,K})\,(D_{R1}+D_{R2}+D_{C2}) + \dots\dots \\ &\quad + P_{1,K}P_{2,K}\dots P_{N,K}(1-P_{N,K})\,(D_{R1}+D_{R2}+\dots+D_{RN}+D_{CN}) + \dots\dots \\ &= \frac{1}{1-P_{1,K}P_{2,K}\cdots P_{N,K}} \left(\sum_{i=1}^{N} \left(\prod_{j=0}^{i-1} P_{j,K} \right) (1-P_{i,K}) \left(\sum_{m=1}^{i} D_{Rm} + D_{Cm} \right) \right) \end{aligned} \tag{6.4}$$

Where, D_{Ri} is an average queueing delay at i-th ring switch for one transmission (i.e. including the case of overflow) and D_{Ci} is an average queueing delay at i-th output port. They are given by M/M/1 formula.

Note that in the homogeneous arrival case (i.e. $\lambda_i = \lambda$, for all i), the equations become simple, and we obtain the polynomial equations in G of order $K+1$.

6.4.3 Numerical Examples and Discussions

In the remaining examples, both simulations and numerical results are presented for comparison purposes. Simulation exactly follows the queueing model in Fig. 6.5, i.e., arrivals of overflowed packets are not assumed to be Poisson process. In plotted performance results, lines and symbols show analytical results and simulations, respectively.

At first, we consider the homogeneous arrival case. Fig. 6.6 illustrates the average switching delay as a function of arrival rate. In the light traffic case, our analysis matches with the simulation quite well. However, for larger values of λ (greater than 0.8), the approximation has a tendency of underestimating delay. We also plot the performance of the unichannel case in which the buffer size is infinite and overflow does not occur. The figure shows that the proposed switch, especially in the case of buffer size $K=$ 3, improves the delay significantly.

Fig. 6.7 shows the expected switching delay as a function of number of buffers. In this figure, ρ is the utilization of output channels. Interestingly, in the light traffic case, the best buffer policy is to have only single buffer (actually, only one, for the job in service). And even in a heavy traffic (ρ = 0.8), the delay is large for $K=1$, reaches a minimum for $K=3$, and then eventually approaches the delay of the unichannel case of 5.3125. In this case, delay is reduced more than 40 %.

Next, we consider the case $\lambda_1 \neq \lambda_2$ and $N=2$. Fig. 6.8 shows the average switching delay as a function of λ_1 for fixed value of λ_2. For buffer $K=5$, D_1 increases slowly with the arrival rate compared with $K=\infty$, and does not saturate when λ_1 is beyond the output channel capacity of its ring μ_C (=1.0). D_2 increases similarly to D_1 for large λ_1. Fig. 6.9 shows the utilizations of output channels for the same parameters in Fig. 6.8. It shows that the proposed switch distributes the packets to the uncongested channels (i.e. channel 2).

Fig. 6.10 shows the average switching delay as a function of the buffer size. For optimal buffer size (i.e. $K=2$), D_1 decreases more than 60% compared with $K=\infty$, while D_2 increases little.

From these results, we can conclude that the proposed switch provides the smaller switching delay and hence more efficient channel utilization.

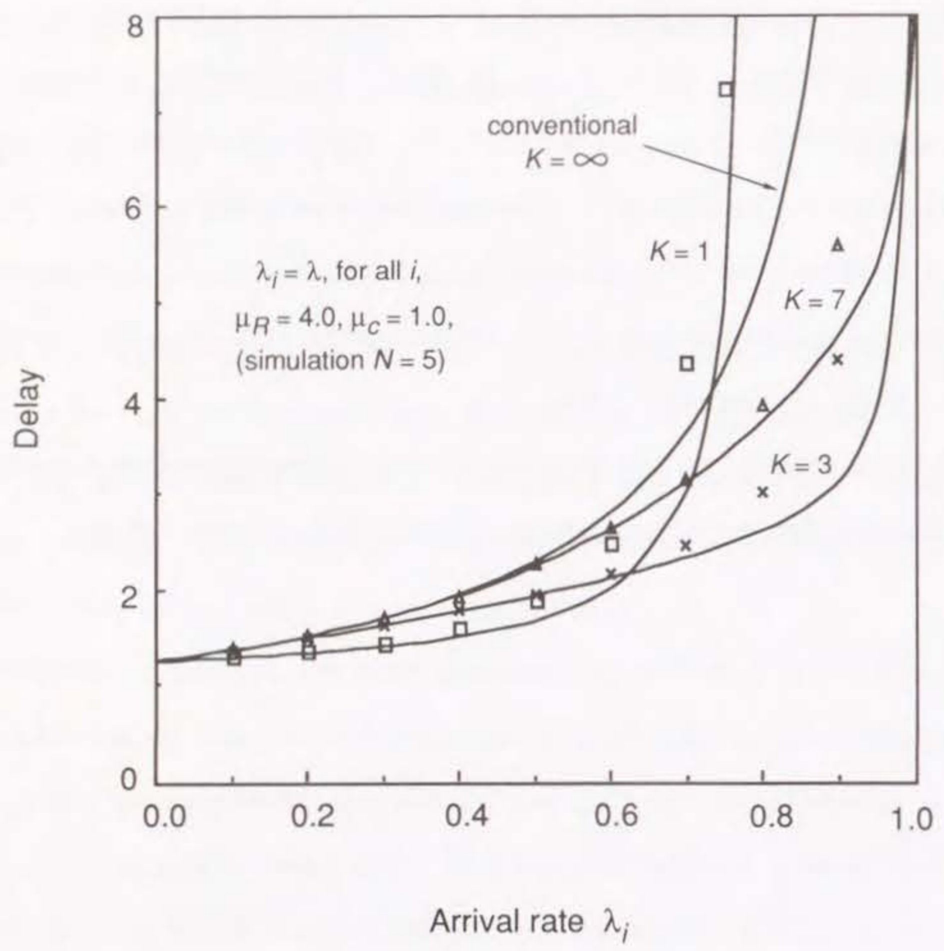


Fig. 6.6 Average switching delay vs. arrival rate.

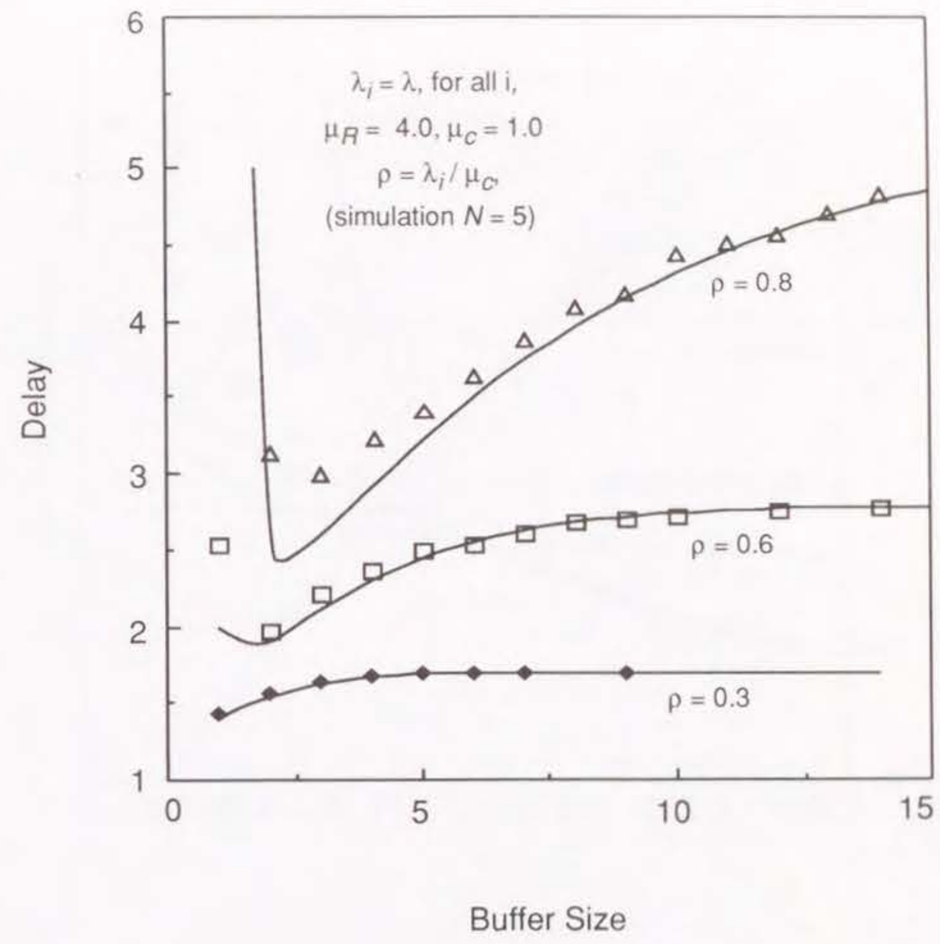


Fig. 6.7 Average switching delay vs. buffer size.

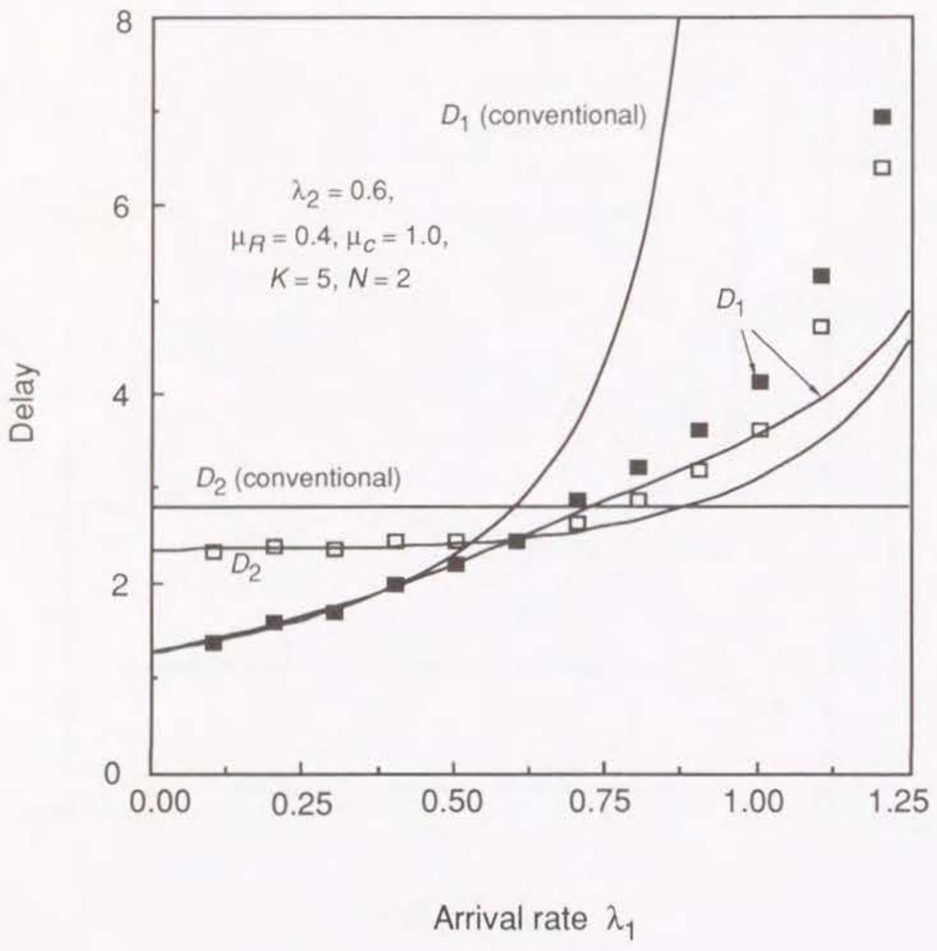


Fig. 6.8 Average switching delay vs. arrival rate for asymmetric arrival model.

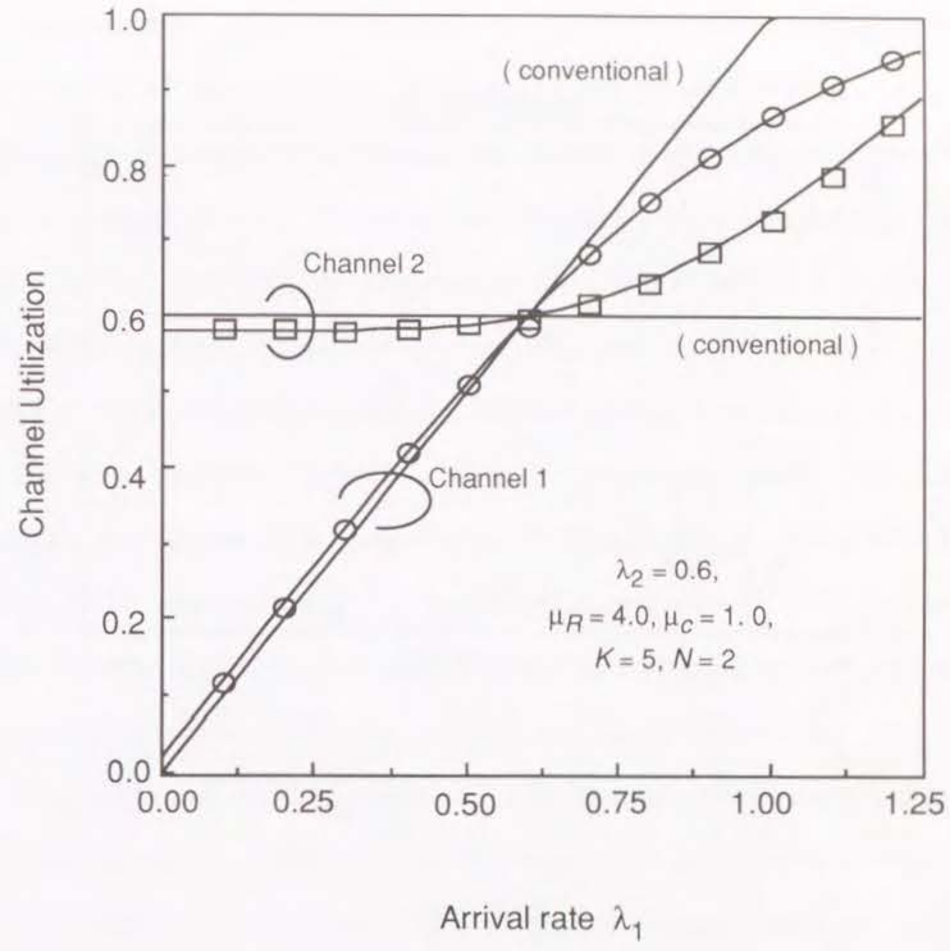


Fig. 6.9 Channel utilization vs. arrival rate for asymmetric arrival model.

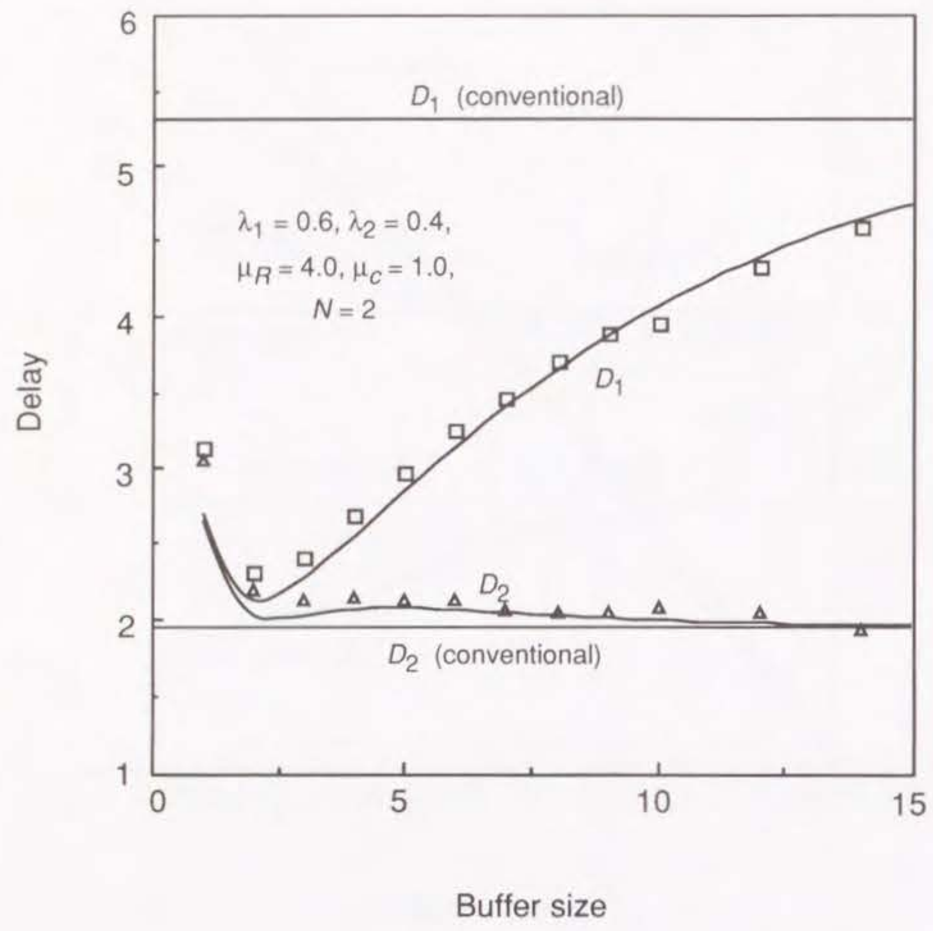


Fig. 6.10 Average switching delay vs. buffer size for asymmetric arrival model.

6.5 Conclusion

In this chapter we have proposed a high-speed packet switch architecture based on a channel group virtual circuit scheme for multimedia communication networks. By allowing packets to select one of several available channels, the switch helps a packet avoid congested output ports and thus provides much lower output queueing delays than conventional single channel switch architectures. The switch accomplished load balancing using a simple, hardware-based, self-routing mechanism, consisting of small number of switching elements, that can handle a flexible length packet.

We also have presented a simple queueing model for preliminary examination of the performance for the proposed switch. The results obtained through both approximate analysis and simulation, have shown that the proposed switch significantly improves channel utilization and switching delay compared with conventional single channel allocation policies.

More detailed examinations for performances of proposed switch remains for the future work. Especially, packet resequencing delay which is required to maintain the packet sequencing and packet lossprobability due to buffer overflow at input port should be evaluated to examine the advantages of the proposed protocol.

CHAPTER 7

Conclusion

In this thesis, we have considered integrated communication systems. The fundamental problem of integration is caused by the differences among various kinds of services both in traffic property and in transmission demand. We focused the attention on the performance evaluation of the effect of new transmission and switching techniques in integrated communication systems, where several performance measures and traffic models were introduced to evaluate individual characteristics of various services.

In Chapter 2, we investigated a voice and data integration on store-and-forward packet switching system. The network was modeled as a type of tandem queueing system with two classes of packets, where a voice packet is lost when the buffer of the switching node is full. We approximately analyzed the end-to-end transmission delay and loss probability of voice packets. The numerical results showed that buffer limit can suppress the traverse time, but that the shorter buffer limit is cut out, the more voice packets are lost.

In Chapter 3, we proposed a hybrid VC-CSMA/CD protocol for CSMA/CD based integrated local area networks. We also analyzed the throughput and delay for connectionless packets and the loss probability of the call for connection oriented packets. The analytical and simulation results showed that the proposed protocol performs well for both types of traffic.

In Chapter 4, simulation models were developed to evaluate performance of packetized voice/data transmission on a token passing ring network. The results suggested that system parameters (e.g., voice packet length, data packet length) can be adjusted to allow token passing ring networks to support both voice and data traffic with acceptable performance.

In Chapter 5, we considered a transportation type bulk arrival bulk service queueing system with composite bulk service discipline as a

fundamental research on future communication networks. The average queue length of customers in the system was approximately derived using probability generating function approach and embedded Markov chain method. Many examples showed that our proposed method provides high approximation accuracy by comparing them with the results of simulations.

In Chapter 6, we proposed a high-speed packet switch architecture based on a channel group virtual circuit scheme for multimedia communication networks. By allowing packets to select one of several available channels, the switch helps a packet avoid congested output ports and thus provides much lower output queueing delays than conventional single channel switch architectures. We also presented a simple queueing model for preliminary examination of the performance for the proposed switch. The results obtained through both approximate analysis and simulation, showed the proposed switch significantly improves channel utilization and switching delay compared with conventional single channel allocation policies.

Direction towards the future research

Since integrated communication systems are still in the phase of rapid growing, the author believes we have a great interest in future researches in this field.

Many future computer users will require world-wide network access. Then all networks including local area networks, special purpose networks and private networks will be connected each other. Interconnection of such networks raises new research problems for investigating the gateway protocols, performance of interconnected integrated systems and so on.

Interconnection of local area networks has given rise to a new network design called metropolitan area network (MAN) which covers a city or a metropolitan area within the range of 10-100 km. Integration onto MAN is one of important research areas in this field.

In the standpoint of the modeling of various kinds of traffic, traffic patterns of applications are depending on the advances of the higher layers such as coding and compressing technique, processing speed of the user's facility, and so on. According as they advance, we have to evaluate the performance of the networks under new traffic models; moreover, we have to develop a new transmission and switching techniques.

Advances of the integrated communication systems may require new performance measures. For example, the multichannel environment as discussed in Chapter 6 raises the problem of packet resequencing. We need a new measure to evaluate this problem.

References

Abbreviations

ICC	International Conference on Communications
IEEE J. Select. Areas Commun.	IEEE Journal of Selected Areas on Communications
IEEE Trans. Commun.	IEEE Transactions on Communications
IEEE Trans. Commun. Theory	IEEE Transactions on Communications Theory
IEEE Trans. Comput.	IEEE Transactions on Computers
IEICEJ	Institute of Electronics, Information and Communication Engineers of Japan
J. Royal Stat. Soc.	Journal of Royal Statistical Society
NTC	National Telecommunications Conference
NCC	National Computer Conference

[BAIL54] N. T. J. Bailey, "On Queueing Process with Bulk Service," J. Royal Stat. Soc., Ser. B 16, pp. 80-87, 1954.

[BHAT64] U. N. Bhat, "On Single-Server Bulk-Queueing Process with Binomial Input," Operations Research, vol. 12, pp. 527-533, 1964.

[BRAD65] P.T. Brady, "A Technique for Investigating On-Off Patterns of Speech," Bell System Technical J., 44, no.1, pp. 1-22. 1965.

[BRAD68] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," Bell System Technical J., vol. 47, no.1, pp. 73-91, 1968.

[BUX81] W. Bux, "Local Area Subnetworks: A Performance Comparison," IEEE Trans. Commun., vol. COM-29, no. 10, 1465-1473, 1981.

[CASN78] S. L. Casner, E. R. Mader and E.R. Cole, "Some Initial Measurements of ARPANET Packet Voice Transmission," NTC Rec., pp. 12.2.1-12.2.5, 1978.

[CCITT85a] CCITT Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," CCITT Red Book, vol. 3, Fascicle 3, 1985.

[CCITT85b] CCITT Recommendation G.72, "32 kbit/s Adaptable Differential Pulse Code Modulation (ADPCM)," CCITT Red Book, vol. 3, Fascicle 3, 1985.

[CCITT85c] CCITT Recommendation J.41, "Characteristics of Equipment for Coding of Analogue High Quality Sound Programme Signals for Transmission on 384 kbit/s Channels," CCITT Red Book, vol. 3, Fascicle 4, 1985.

[CHLA85] I. Chlamtac, "An Ethernet Compatible Protocol for Real-Time Voice/Data Integration," Computer Networks and ISDN Systems, vol. 10, pp. 81-96, 1985.

[CHLA88] I. Chlamtac and A. Ganz, "Channel Allocation Protocols in Frequency-Time Controlled High-Speed Networks," IEEE Trans. Commun., vol.36, no. 4, pp. 430-440, 1988.

[COHE81] D. Cohen, "Packet Communication of Online Speech," Proc. NCC, pp. 169-176, 1981.

[COHE69] J. W. Cohen, *The Single Server Queue*, Wiley: New York, 1969.

[COVI77] G.J. Coviello, O. L. Lake and G. R. Redinbo, "System Design Implications of Packetized Voice," Proc. ICC, pp. 38.2-49 - 38.2-53, 1977.

[DeTR84] J. D. DeTreville, "A Simulation-Based Comparison of Voice Transmission on CSMA/CD Networks and on Token Buses," AT&T Bell Lab. Tech. J., vol. 63, no. 1, pp. 33-55, 1984.

[FALC85] R. M. Falconer and J. L. Adams, "Orwell: A Protocol for an Integrated Services Local Network," British Telcom. Technol. J., vol. 3, no. 4, pp. 27-35, 1985.

[FORG75] J. W. Forie, "Speech Transmission in Packet-Switched Store-and-Forward Networks," Proc. NCC, pp. 137-142, 1975.

[FORG77] J. W. Forgie and A. Nemeth, "An Efficient Packetized Voice/Data Network Using Statistical Flow Control," Proc. IEEE ICC, pp. 38.2-44 -38.2-48, 1977.

[GITM77] I. Gitman, H. Frank, B. Occhiogrosso, and W. Hsieh, "Issues in Integrated Network Design," Proc. IEEE ICC, pp. 38.1-36 - 38.1-43, 1977.

[GITM78] I. Gitman and H. Frank, "Economic Analysis of Integrated Voice and Data Networks: A Case Study," Proc. IEEE, vol. 66, no. 11, pp. 1549-1570, 1978.

[GONS82] T. A. Gonsalves, "Packet-Voice Communication on an Ethernet Local Computer Network: An Experimental Study," Xerox Palo Alto Research Center Tech. Rep. CSL-82-5, 1982.

[GRUB81] J. G. Gruber, "Delay Related Issues in Integrated Voice and Data Networks," IEEE Trans. Commun., vol. COM-29, no. 6, pp. 786-800, 1981.

[GRUB82a] J. G. Gruber, "Performance Considerations for Integrated Voice and Data Networks," Computer Communications, vol. 4, no. 3, pp. 106-126, 1982.

[GRUB82b] J. G. Gruber, "A Comparison of Measured and Calculated Speech Temporal Parameters Relevant to Speech Activity Detection," IEEE Trans. Commun., vol. COM-30, no. 4, 728-738, 1982.

[HASE64] T. Hasegawa, Y. Tezuka and Y. Kasahara, "Digital Data Dynamic Transmission Systems," IEEE Trans. Commun. Theory, vol. COM-12, no. 1, pp. 58-65, 1964.

[HASE66] T. Hasegawa, Y. Tezuka and Y. Kasahara, "Transmission Delay and Channel Loading in Digital Data Dynamic Transmission Systems," IEEE Trans. Commun. Theory, vol. COM-14, no. 2, pp. 94-101, 1966.

[HUI87] J. Hui, "A Broadband Packet Switch for Multi-Rate Services," Proc. IEEE ICC, pp. 782-788, 1987.

[IEEE85a] IEEE Standards 802.3-1985, "Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications," IEEE Standards for Local Area Networks, 1985.

[IEEE85b] IEEE Standards 802.5-1985, "Token Ring Access Method and Physical Layer Specifications," IEEE Standards for Local Area Networks, 1985.

[JAIS60] N. K. Jaiswal, "Time-Dependent Solution of the Bulk-Service Queueing Problem," Operations Research, vol. 8, pp.773-781, 1960.

[JOHN81] D. H. Johnson and G. C. O'Leary, "A Local Access Network for Packetized Digital Voice Communication," IEEE Trans. Commun., vol. COM-29, no. 5, pp. 679-688, 1981.

[KISH89] F. Kishono, K. Manabe, Y. Hayashi and H.Yasuda, "Variable Bit-Rate Coding of Video Signals for ATM Networks," IEEE J. Select. Areas Commun., vol. SAC-7, no. 5, pp. 801-806, 1989.

[KLEI64] L. Kleinrock, *Communication Nets,* Mcgraw-Hill Book Co., pp. 49-56, 1964.

[LAM76] S. S. Lam, "Store-and-Forward Buffer Requirements in a Packet Switching Network," IEEE Trans. Commun., vol. COM-24, no. 4, pp. 394-403, 1976.

[MATS88] Y. Matsumoto, Y. Takahashi and T. Hasegawa, "Performance Comparison of Two CSMA/CD Systems with Prioritized Retransmission," Trans. IEICEJ, vol. J71-B, no. 6, pp. 678-689, 1988 (in Japanese).

[MAXE82] N. F. Maxemchuk, "A Version on CSMA/CD That Yields Movable TDM Slots in Integrated Voice/Data Local Networks," Bell System Technical J., vol. 61, no. 7, pp. 1527-1550, 1982.

[MEDI85] J. S. Meditch and Y. Zhao, "Framed TDMA/CSMA for Integrated Voice-Data Local Area Networks," Proc. INFOCOM, pp. 10-17, 1985.

[MILL59] R. G. Miller, "A Contribution to the Theory of Bulk Queues," J. Royal Stat. Soc., Ser. B 21, pp.320-337, 1959.

[MINO79] D. Minoli, "Issues in Packet Voice Communication," Proc. IEE, vol. 126, no. 8, pp. 729-740, 1979.

[MOWA80] O. A. Mowafi and W. J. Kelly, Jr., "Integrated Voice/Data Switching Techniques for Future Military Networks," IEEE Trans. Commun., vol. COM-28, no. 9, pp. 1655-1662, 1980.

[NEUT79] M.F. Neuts, "Queues Solvable without Rouche's Theorem," Operations Research, vol. 27, pp. 767-781, 1979.

[NOJI87] S. Nojima, E. Tsutsui, H. Fukuda, and M. Hashimoto, "Integrated Services Packet Network Using Bus Matrix Switch," IEEE J. Select. Areas Commun., vol.SAC-5, no.8, pp.1284-1292, 1987.

[NUTT82] G.J. Nutt and D. L. Bayer, "Performance of CSMA/CD Networks Under Combined Voice and Data Loads,," IEEE Trans. Commun., vol. COM-30, no.1, pp. 6-11, 1982.

[OHTS84] K. Ohtsuki, Y. Takahashi and T. Hasegawa, "Analysis for Traverse Time in an Integrated Communication System," Proc. IEEE ICC, pp. 1276-1279, 1984 (also in *Links for the Future*, North Holland, 1984).

[OHTS85] K. Ohtsuki, T. Shimizu and H. Hasegawa, "Approximate Analysis for Bulk Queueing System with Composite Service Discipline," Proc. of International Teletraffic Congress, pp. 3.2A-2-1 - 3.2A-2-6, 1985.

[OHTS89] K. Ohtsuki, Y. Takahashi and T. Hasegawa, "Performance Analysis of a Hybrid VC-CSMA/CD Protocol in Integrated Local Area Bus Networks," Trans. IEICEJ, vol. J72-B, no. 9, pp. 710-720, 1989 (in Japanese).

[OHTS90] K. Ohtsuki, H. Okada and H. Hasegawa, "A Packet Switch Architecture with Multiple Virtual Circuit Scheme in High-Speed Communications Networks," IEICEJ Technical Report, vol. IN90-6, pp. 7-12, 1990 (in Japanese).

[OHTS91a] K. Ohtsuki, H. Okada and T. Hasegawa, "A Packet Switch Architecture with Channel Group Virtual Circuit Scheme in High-Speed Communication Networks," Trans. IEICEJ, vol. J74-B, no. 1, pp. 20-27, 1991 (in Japanese).

[OHTS91b] K. Ohtsuki, K. Takemura, J. F. Kurose, H. Okada and Y. Tezuka, "A High-Speed Packet Switch Architecture with a Multichannel Bandwidth Allocation," to appear in IEEE INFOCOM, 1991.

[OKAD84] H. Okada, "CSMA/CD/MPD Access Control Method in Bus-Type Local Area Networks for Integrated Data and Voice," Trans. IEICEJ, vol. J67-D, no. 1, pp. 117-124, 1984 (in Japanese).

[PATT88] A. Pattavina, "Multichannel Bandwidth Allocation in a Broadband Packet Switch," IEEE J. on Selected Areas on Commun., vol. SAC-6, no. 9, pp. 1489-99, 1988.

[RABI78] L. R. Rabiner and R. W. Shcafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ : Prentice-Hall, 1978.

[RIOS85] M. Rios and N. D. Georganas, "A Hybrid Multiple-Access Protocol for Data and Voice-Packet over Local Area Networks," IEEE Trans. Comput., vol. C-34, no. 1, pp. 90-94, 1985.

[ROSS86] F. E. Ross, "FDDI : a Tutorial," IEEE Communications Magazine, vol. 24, no.5, 1986.

[SCHW76] L. Schweizer, "Performance of Terrestrial and Satellite 64 kbit/s Paths : Requirements of Voice and Data, and Standards of the

Future Integrated Digital Network (ISDN)," Proc. IEEE ICC, pp. A1.4.1-A1.4.5, 1983.

[SEN87] P. Sen, "Modelling and Performance Analysis of a Hybrid Protocol for Voice-Data Integration in Local Area Networks," Proc. of INFOCOM, pp. 981-986, 1987.

[SUDA85] T. Suda, C. Yuen and K. Ohtsuki, "Performance Evaluation of Packetized Voice Transmission on a Token Passing Ring Network," Proc. IEEE GLOBECOM, pp. 17.3.1 - 17.3.5, 1985.

[TAKE87] T. Takeuchi, T. Yamaguchi, H. Niwa, H. Suzuki and S. Hayano, "Synchronous Composite Packet Switching : A Switching Architecture for Broadband ISDN," IEEE J. Select. Areas Commun., vol. SAC-5, no. 5, pp. 1365-1376, 1987.

[TOBA80] F. A. Tobagi, and V. B. Hunt "Performance Analysis of Carrier Sense Multiple Access with Collision Detection," Computer Networks, vol. 4, pp. 245-259, 1980.

[TOBA82] F. A. Tobagi, "Carrier Sense Multiple Access with Message-Based Priority Functions," IEEE Trans. Commun., vol. COM-30, no. 1, pp. 185-200, 1982.

[TURN86] J. S. Turner, "Design of an Integrated Services Packet Network," IEEE J. Select. Areas Commun., vol. SAC-4, no. 8, pp. 1373-1379, 1986.

[WATA83] M. Watanabe, T. Shimizu, Y. Takahashi and T. Hasegawa, "Approximate Analysis for Exclusive Bulk Service Queues," Kokyuroku, 490, Research Inst. Math. Science, Kyoto University, pp. 227-247, 1983 (in Japanese).

[WHIT75] J. A. White, J. W. Schmidt and G. K. Bennett, *Analysis of Queueing Systems*, Academic Press, 1975.

[YEH87] Y.-S. Yeh, M. G. Hluchyj, and A.S. Acampora, "Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," IEEE J. Select. Areas Commun., vol. SAC-5, no. 8, pp. 1274-1283, 1987.